The Preserve: Lehigh Library Digital Collections

Leveraging Data Science to Aid Clinicians in Understanding Bone Marrow Aspirate Concentrate as a Therapeutic

Citation

Herna, Colin. *Leveraging Data Science to Aid Clinicians in Understanding Bone Marrow Aspirate Concentrate As a Therapeutic*. 2026, https://preserve.lehigh.edu/lehigh-scholarship/graduate-publications-theses-dissertations/theses-dissertations/leveraging-data.

Find more at https://preserve.lehigh.edu/

This document is brought to you for free and open access by Lehigh Preserve. It has been accepted for inclusion by an authorized administrator of Lehigh Preserve. For more information, please contact preserve@lehigh.edu.

Leveraging Data Science to Aid Clinicians in Understanding Bone Marrow Aspirate Concentrate as a Therapeutic

by

Colin Herna

A Dissertation

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Bioengineering

Lehigh University

January 2026

© 2025 Copyright
Colin Herna

Dissertation Signature Sheet

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Colin Herna

Leveraging Data Science to Aid Clinicians in Understanding Bone Marrow Aspirate Concentrate as a Therapeutic

_____

Defense Date

_____

Dr. Sabrina Jedlicka

_____

Approved Date

Committee Members:

_____

Dr. Sabrina Jedlicka

_____

Dr. Susan Perry

_____

Dr. Xuanhong Cheng

_____

Dr. Ranjan Sachdev

ACKNOWLEDGMENTS

“A frog in a well knows nothing of the great sea” – To the people who have served as my guides, both big and small, I am grateful. While it is only my name listed on this document, the accumulation of my efforts is and was only ever possible because they were scaffolded by those around me. For that I do not believe words will really convey my thanks, however I hope that when those people look at me, they feel as though their efforts were not wasted. Thank you.

Table of Contents

List of Tables

List of Figures

ABSTRACT

Osteoarthritis (OA) represents one of the most pervasive degenerative joint diseases worldwide, characterized by pain, swelling, and progressive functional decline. Current therapeutic approaches largely focus on symptom management rather than the underlying biological mechanisms driving tissue degeneration. Within this therapeutic landscape, bone marrow aspirate concentrate (BMAC) has emerged as a promising orthobiologic, capable of leveraging the body's own regenerative and anti-inflammatory potential. Composed of platelets, growth factors, cytokines, and mesenchymal stem cells (MSCs), BMAC functions as a point-of-care biologic that simultaneously delivers multiple therapeutic agents within a single injectable. Despite its widespread clinical use, however, the proteomic underpinnings that define BMAC's therapeutic efficacy remain poorly characterized within the literature.

This project seeks to address that gap by combining the principles of proteomics and data science to explore BMAC as both a biological and computational system. From a biological standpoint, the composition of BMAC varies according to patient specific factors such as age, sex, body mass index (BMI), smoking status, and the anatomical site of harvest in addition to procedural differences. From a computational standpoint, these variations manifest as a high-dimensional dataset in which biological patterns are often non-linear, multivariate, and subtle. The central premise of this work is therefore that data-driven modeling provides a necessary and scalable framework to

uncover the latent biological structure that traditional univariate analyses often struggle to capture.

The long-term goal of this project is to establish a proteomic and analytical framework that enables clinicians and researchers to characterize BMAC in a reproducible, quantitative, and accessible manner. The short-term objective is to examine whether BMAC samples derived from different anatomical regions, specifically the iliac crest and humeral head, exhibit distinct proteomic profiles that are further modulated by patient demographics. To test this hypothesis, BMAC samples were analyzed using high-throughput antibody microarray technology, allowing for semi-quantitative measurement of 109 biomarkers across cytokines, chemokines, and growth factors. The dataset produced from these analyses forms the foundation for downstream inferential statistics and machine learning.

Within this framework, probabilistic reasoning, statistical inference, and machine learning converge to create a rigorous and reproducible method for disentangling variability within patient-derived biologics. The application of these techniques enables the identification of latent protein networks and interaction signatures that might otherwise remain obscured in conventional univariate analyses. These computational approaches do not replace biological interpretation but rather extend it, providing a scaffold upon which biological meaning can be quantified and compared across diverse patient populations and clinical contexts.

The implications of this work are therefore twofold. Biologically, it provides one of the most comprehensive proteomic characterizations of BMAC to date, revealing how anatomical source and patient demographics contribute to the underlying protein composition of this therapeutic.

Ultimately, this work seeks to make the study of BMAC transparent, quantitative, and clinically meaningful. By embedding the analysis of biologics within a data science framework, this work contributes to the growing field of personalized regenerative medicine.

Chapter 1:

Introduction to Bone Marrow Aspirate Concentrate

1.1 Bone Marrow Aspirate Concentrate Overview

Osteoarthritis (OA), is defined by the World Health Organization as a degenerative joint condition characterized by pain, swelling, and stiffness; the entire joint and its surrounding tissues are often affected. OA predominately affects the knees, followed by the hips, spine, and hands.[1] While it is most prevalent in adults over the age of 55, the pathology often manifests as early as the late 40's to mid-50's, resulting in progressive pain and functional decline.[2] Current treatment options remain limited and usually focus on symptom management rather than the underlying pathology.

In this therapeutic landscape, bone marrow aspirate concentrate (BMAC) is what clinicians may describe as having incredible therapeutic potential. Best described by Lisa Fortier, BMAC is, "a point of care treatment which both uniquely and simultaneously delivers growth factors, anti-inflammatory proteins, and mesenchymal stem cells in a single-injectable orthobiologic".[3] This combination of constituents is what makes BMAC uniquely suited to target not just the core pathologies of OA, but also tackle the broader landscape of cartilage repair across many other forms of musculoskeletal defects. BMAC's

therapeutic potential is rooted in several intrinsic properties: it is an autologous, immunologically safe product that can deliver innate anti-inflammatory molecules like interleukin-1 receptor antagonist (IL-1Ra), a plethora of tissue-stimulating growth factors, and mesenchymal stem cells (MSCs) with their own profound immunomodulatory and regenerative properties. Within the expanding field of orthobiologics, BMAC represents an excellent therapeutic that leverages the body's own regenerative and anti-inflammatory machinery.

1.2 Synergistic Characteristics of Bone Marrow Aspirate Concentrate

BMACs unique regenerative properties are the result of several key features such as platelets, anti-inflammatory molecules and growth factors, in conjunction with mesenchymal stem cells work together to form a sum which is greater than its parts, in this section we will take a deeper look into what each of these components contributes (Fig 1.1):

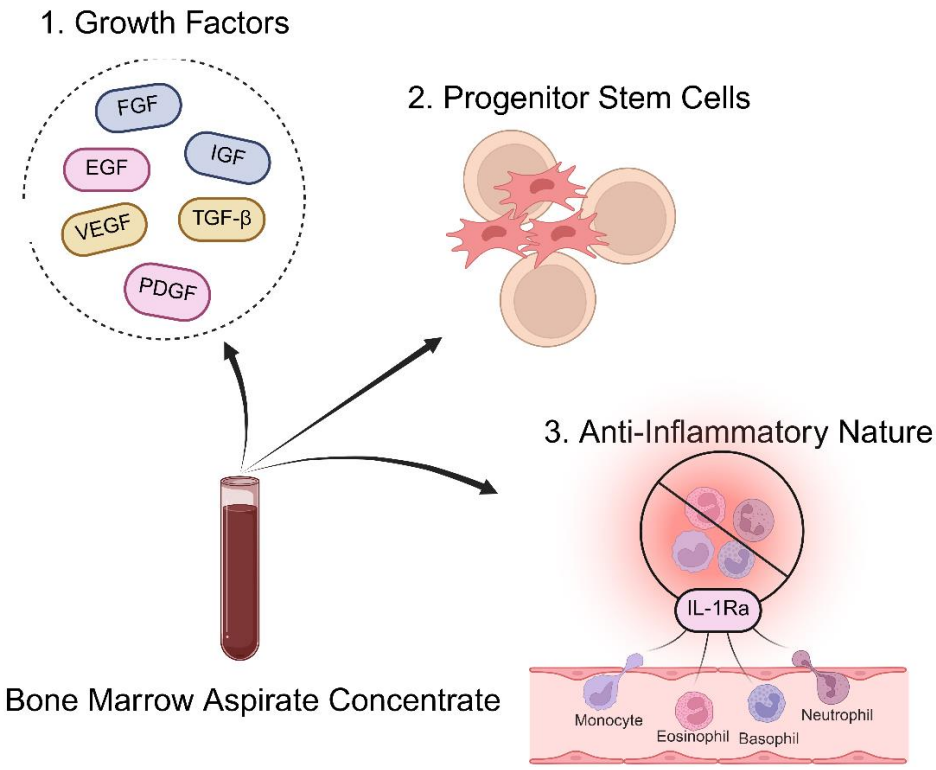

Figure 1.1 Properties of BMAC: *Key characteristics of BMAC as a therapeutic include growth factors, progenitor cells, and anti-inflammatory molecules. Created in BioRender. Jedlicka, S. (2025) https://BioRender.com/d69r683.*

1.2.1 Anti-inflammatory Properties

The interleukin-1 family contains three members, IL-1α, IL-1β, and IL-1Ra; IL-1 is described to have an outstanding reach, affecting almost every type of cell, and done so in conjunction with other cytokines and mediator molecules. IL-1 is regarded as a highly inflammatory cytokine whose range between toxic and therapeutic is narrow at best. Fortunately, there are a series of innate biological regulation processes which seek to keep the proinflammatory IL-1α and β in check. The first major control is the presence of two receptors (IL-1R1 or IL-1R2) which can accept IL-1α and β signal molecules; despite two receptors, only IL-1R1 will elicit a downstream signal cascade. IL-1R2 simply acts as a decoy due

to lacking the cytosolic component required for signal transduction, thereby diluting the pool of available signal molecules. The second major control comes from interleukin-1 receptor antagonist (IL-1ra, IRAP). IRAP acts as a competitive inhibitor which attempts to outcompete IL-1α and β for available IL-1R1 binding domains.[4–6] This signal pathway has been summarized in Fig 1.2 modified from *Kaneko et al*.[7]

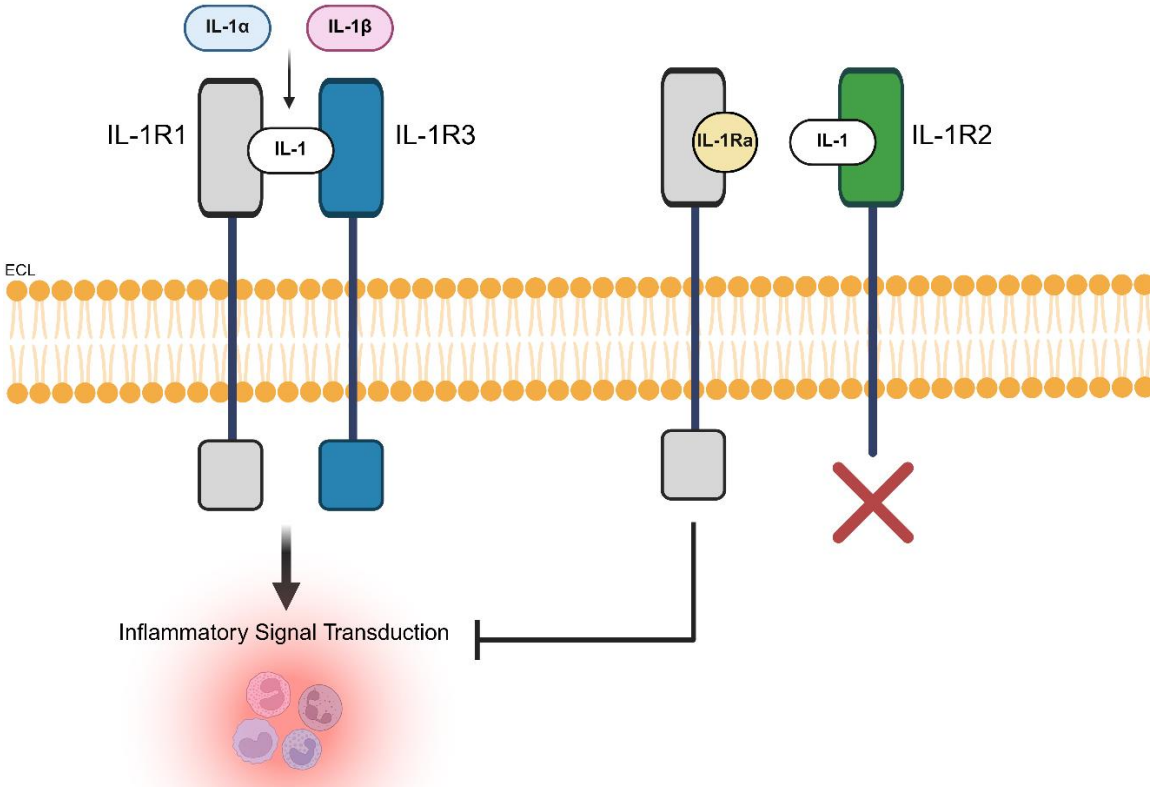

Figure 1.2 IL-1 Receptor Mechanism: *Interleukin-1 receptors and inhibitors for IL-1 signaling. Il-1R1 and its co-receptor IL1-R3 (IL-1RAcP) interact with both IL-1α and β for signal transduction. IL-1Ra can bind solely to IL1-R1 to inhibit IL-1 signaling. IL-1R2 is a decoy receptor that lacks the cytosolic component for signal transduction.[7] Extracellular Leaflet (ECL). Created in BioRender. Jedlicka, S. (2025) https://BioRender.com/c0lzlnt*

The anti-inflammatory properties of BMA and BMAC can be greatly attributed to the innately high presence of interleukin-1 receptor antagonist (IL-1ra, IRAP). IRAP's anti-inflammatory properties are so notable, that there are several commercial systems available to essentially farm IRAP. This process

typically involves incubating blood samples for 8 to 24 hours to induce enhanced IRAP production in the local monocyte population. These kits have shown promise as far as reducing the Visual Analogue Scale (VAS) and the Western Ontario and McMaster Universities Arthritis Index (WOMAC) scores of OA patients whose knees were injected with the isolated IRAP protein. Despite the promise of these kits, they remain unapproved as a treatment by the Food and Drug Administration (FDA) because they are considered to manipulate blood outside of the body.[3]

Interestingly, the native concentrations of IRAP found in both BMA and BMAC are an order of magnitude greater than what is obtainable via these kits. Specifically, the amount of the IRAP these kits can produce is only a fraction of what is intrinsically found in bone marrow. Consequently, both BMA and BMAC represent exceptionally potent sources of the anti-inflammatory IRAP protein.

1.2.2 Platelets and Growth Factor Delivery in BMAC

Platelets are anucleated cytoplasmic fragments derived from megakaryocytes that play essential roles in hemostasis, inflammation, and tissue regeneration. Within BMAC, they act as mobile storage and delivery units for a myriad of growth factors, cytokines, and chemotactic molecules. Upon activation, platelets release α-granule contents containing transforming growth factor-β (TGF-β), insulin-like growth factor (IGF), platelet-derived growth factor (PDGF), fibroblast growth factor (FGF), vascular endothelial growth factor (VEGF), and epidermal growth factor (EGF), among others.[8,9]

These signaling molecules orchestrate tissue repair by stimulating angiogenesis, fibroblast proliferation, and extracellular matrix remodeling, while also attracting mesenchymal progenitor cells that can differentiate into osteoblasts, chondrocytes, and fibroblasts. Compared with platelet-rich plasma (an alternative orthobiologic), BMAC exhibits similar platelet concentrations but contains a higher leukocyte fraction, which may further support a regenerative microenvironment through immune modulation rather than excess inflammation.[3]

The growth factors released from platelet α-granules and other marrow constituents exert diverse effects ranging from tissue repair and maintenance to immune modulation and cell-cycle regulation. The key growth factors include platelet-derived growth factor (PDGF), vascular endothelial growth factor (VEGF), transforming growth factor-β (TGF-β), fibroblast growth factor (FGF), insulin-like growth factor (IGF), and epidermal growth factor (EGF).3 Their primary biological roles are summarized as follows:

PDGF – The primary mitogen (cell division signal) for connective tissue cells; PDGF is released by platelets at the site of an injury, resulting in fibroblasts and smooth muscle cells to migrate and populate the injury site before proliferating and producing new matrix for tissue repair.[10,11]

VEGF – The master regulator for angiogenesis (new blood vessel formation). VEGF stimulates the cells lining the blood vessels to initiate proliferation and migration, while also increasing the vascular permeability of the existing blood vessels to increase the production of new blood vessels.[12]

TGF-β – Key modulator of cell proliferation, metabolism, and differentiation. TGF-β has the capacity to prevent autoimmune responses by suppressing immune cell activation. It is also able to act as a tumor suppressor to inhibit cell division and during embryonic development it is responsible for promoting the epithelial to mesenchymal transition.[13]

FGF – Typically involved in wound healing and development; FGF works in tandem with other growth factors such as VEGF and TGF-β. [14,15]

IGF – Primary mediator of growth hormone production which promotes systemic growth of the entire musculoskeletal system. IGF is also involved in the modulation of cell metabolism, survival and proliferation.[16,17]

EGF – Stimulates the differentiation, chemotaxis, and proliferation of epithelial and epidermal cells.[18]

1.2.3 Mesenchymal Stem Cells

In addition to these growth factors, BMAC also houses Mesenchymal stem cells (MSCs) which have been shown in a varying capacity to have major effects on chondrogenesis, chondroprotection, and immune modulation in arthritic joints, notably these are not found in platelet rich plasma which is a common alternative to BMAC.[3] The MSC component of BMAC has long been thought to be the primary driver for the regenerative potential of BMAC. Many early bodies of work saw BMAC as a biological augment whose main goal was to harvest these unspecialized cells from the bone marrow, concentrate them, and reintroduce them to the site of an injury to promote healing; this in large part came from

MSCs ability to differentiate into cartilage, tendons, bones, and muscles, among many other notable properties.[8,19]

MSCs are multipotent stromal cells first described in the 1960s. They are characterized by their capacity for self-renewal and differentiation into mesodermal lineages such as osteoblasts, adipocytes, and chondroblasts.[20–24] Although they represent a small fraction (0.001% to 0.01%) of the mononuclear cells in bone marrow, their therapeutic potential has made them a focal point in regenerative medicine.[19,23,25–27] The Mesenchymal and Tissue Stem Cell Committee of the International Society for Cellular Therapy has established minimal criteria to define MSCs: they must be plastic-adherent under standard culture conditions, express CD105, CD73, and CD90 while lacking expression of hematopoietic markers (CD45, CD34, etc.) and HLA-DR; and must demonstrate in vitro differentiation into osteoblasts, adipocytes, and chondroblasts.[28]

The therapeutic profile of MSCs is attributed to a combination of remarkable properties. Key features include intrinsic tropism, or a homing mechanism, enabling them to mobilize to sites of inflammation.[19] Upon arrival, they exert potent anti-inflammatory and immunomodulatory effects by upregulating anti-inflammatory cytokines (e.g., IL-1RA) and downregulating pro-inflammatory ones, while also modulating the activity of B and T lymphocytes, natural killer cells, and macrophages.[19,21,29,30] Furthermore, MSCs promote tissue repair and regeneration through the regulation of cell proliferation, antioxidant activity, and differentiation (Fig 1.3).[31,32] While these effects were initially attributed to direct differentiation, it is now recognized that paracrine signaling is

the primary mechanism.[19,27] Rather than differentiating in mass, MSCs release a cocktail of bioactive factors—including proteins, mRNAs, and miRNAs—packaged within exosomes and small extracellular vesicles (sEVs).[33,34] This collective set of secreted factors, known as the secretome, is the principal mediator of MSC therapeutic effects.[30,35,36]

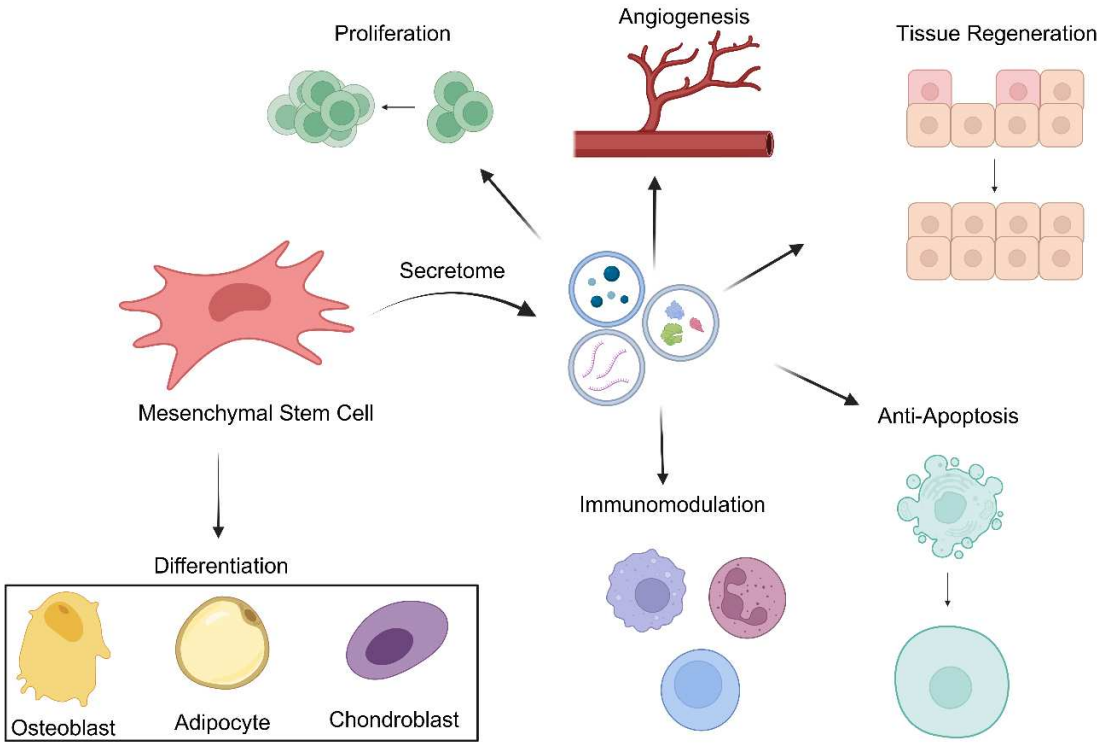

Figure 1.3 Properties of Mesenchymal Stem Cells: *Illustration of mesenchymal stem cells (MSC) generalized functions. MSCs are multi-potent stem cells capable of differentiating into osteoblasts, adipocytes, and chondroblasts. Through paracrine signaling the secretome is secreted into the local environment which stimulates a host of effects including cell and tissue regeneration, immunomodulation, rescuing cells from apoptosis, and stimulating new blood vessel formation. Created in BioRender. Jedlicka, S. (2025) https://BioRender.com/u1yjbfg.*

A critical aspect of the secretome is that its composition is not fixed; it is dynamically influenced by environmental conditions. [30,35,36] This variability introduces a question: what factors modulate the secretome, and how can they

be leveraged to maximize therapeutic efficacy? While numerous donor-specific variables have been investigated—including age, sex, BMI, smoking status, and tissue source—the impact of these factors remains an area of active research, with studies often reporting mixed conclusions.[37–42] For instance, some evidence suggests that advancing donor age can diminish MSC functionality, highlighting the need to systematically isolate and understand these demographic variables.[37–39] Therefore, a key challenge in the field is to precisely define how specific factors alter the MSC secretome to ensure consistent and potent clinical outcomes.

1.3 Obtaining BMAC

Due to the lack of standardization around how BMAC is procured, there are many different methods of obtaining BMAC; however, they generally follow this protocol: Sedate or anesthetize the patient, identify the bone marrow extraction site, harvest the bone marrow aspirate, then centrifuge the aspirate into a concentrated mixture.

Gowd et al. describes their process as putting the patient under a general anesthesia in a supine position as opposed to more conventional beach chair position and can be seen in Fig 1.4. In contrast, authors Chahal et al. elect for a prone position with light sedation. In both cases, what follows is the identification of the anatomical landmarks near the harvest site prior to sterilization and draping. At this point one of many commercially available bone marrow extraction kits (e.g. Arthrex or Biomet Biologics) are used to penetrate the harvest site, apply an anticoagulant, and begin harvesting the marrow.[8,43]

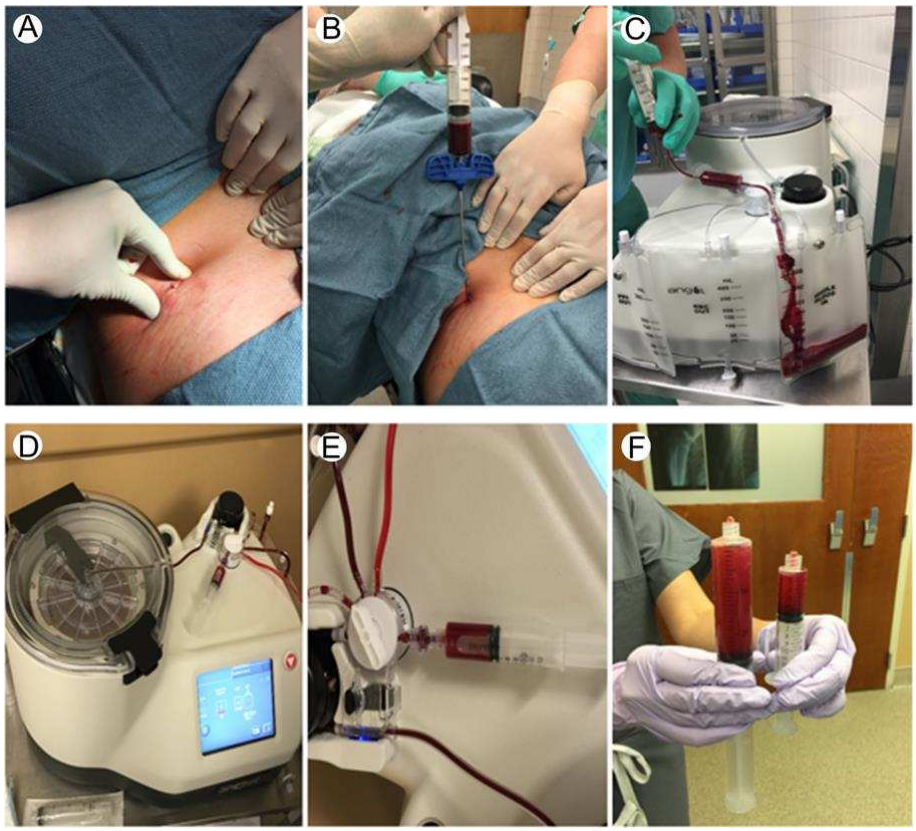

Figure 1.4 Acquisition of BMAC: *Depiction of acquiring BMAC from Gowd et al.[8] (A) Identifying anterior superior iliac spine. (B) Aspiration of bone marrow. (C) Loading marrow into Arthrex Angel system. (D) Arthrex Angel system processing aspiration sample. (E) Production of bone marrow aspirate concentrate. (F) Comparison of unconcentrated bone marrow (left) to BMAC (right).*

The aspiration can be done all at once or in aliquots depending on factors such as the harvest site, or desired patient outcome; however, the literature suggests that ~60 mL of total volume is desired for the subsequent concentration procedure.[3] The most common method for processing the BMA (45-60 mL) into BMAC (3-8 mL) is centrifugation. This process separates the sample into distinct layers based on cellular density and sedimentation rates: one enriched with leukocytes (white blood cells), mesenchymal stem cells (MSCs), hematopoietic stem cells (HSCs), and platelets, and another comprised primarily of red blood cells (RBCs). The nucleated cell–rich layer is retained and resuspended in the platelet-poor plasma fraction, yielding the final BMAC product.[3]

Beyond the procedural variability, there is also a significant disparity between harvest site selection. The overwhelming majority of studies describe aspirating marrow from the iliac crest, this is likely due to be the result of the extractable volume of bone marrow, total progenitor cell yield, and anecdotal evidence suggesting preferred clinical patient outcomes.[3,44,45] A particularly influential study by Hyer et al. compared osteoblastic progenitor cell yield from the iliac crest, proximal aspect of the tibia, and calcaneus. Their results suggested that while all sites are viable, the progenitor cell yield from the iliac crest was the greatest and therefore quantitatively the best option.[41]

Despite the assessment of Hyer et al, it is important to note that there is a considerable amount of ongoing research around BMAC and more specifically the total composition of BMAC, not just the cellular aspect of the injectable. Progenitor cell yields are not the defining factor of BMAC, it is important to keep in mind that BMAC is a product greater than the sum of its parts. Furthermore, there are more niche instances of BMAC extraction from locations like the humeral head which become accessible during a full thickness rotator cuff tear; the access to such a harvest site then enables clinicians to both perform a repair and marrow harvest in the same site rather than pivoting from the harvest site to the surgical site.[46]

1.4 Clinical Outcomes Using BMAC

The clinical outcomes reported for BMAC treatments often report pain relief, functional improvement, and structural pair for musculoskeletal defects. However, the evidence remains variable in terms of its quality, with a majority of

the studies having small sample sizes, short follow-up periods, or heterogeny with respect to the BMAC product produced, dosage administered, and alignment of reported outcomes.

Recent examples shown BMAC's effectiveness in knee OA: A cohort of 213 patients (285 knees) with Kellgren-Lawrence grades II-III OA treated by BMAC derived from the anterior iliac crest with a single spin centrifugation, demonstrated large and statistically significant improvements in pain and function over a 7-month period. The visual analog scale (VAS) pain dropped from ~4.3 to ~0.4 and the Knee Society knee score and function score improved in a similar capacity. The authors of this study deemed BMAC to be a reliable and effective treatment.[47] Another study, used cohort of 17 knee OA patients underwent, BMAC treatment using a fenestrated trocar system (no centrifugation) between March 2018 and 2019. 10 mL of BMAC was harvested and injected into the affected knees of each patient. The clinical outcomes were monitored at 0, 6, and 12 weeks. The results showed significant improvements in all scores over the 12 weeks; specifically the Knee Injury and OA Outcome Score (KOOS) improved day to day activity by ~28 points, quality of life by ~33 points, sport and recreation by ~20 points, and pain by ~30 points, all of which were measured against a 100 point scale. Lysholm scores improved by ~22 points and VAS scores decreased by ~3.3 points.[48]

In the context of augmenting existing repairs, BMAC has been used to aid in tendon-bone healing and rotator cuff tear repair. Several studies reported by A.K.Gowd et al.[8] suggest that when BMAC is used in tandem rotator cuff repair

(RCR), tendon integrity as assessed by MRI or imaging was better preserved with functionality and strength sores such as the UCLA shoulder score improving over a 12 month period.[46,49,50] When used in hip arthroscopy, patients who received the BMAC augmentation for cartilage damage during labral repair saw a significantly greater functional outcome improvements at 12 and 24 months compared to the patients without BMAC. Additionally, for full-thickness chondral flaps, BMAC outperformed microfracture in reaching the minimal clinically important difference (MCID) for International Hip Outcome Tool-33 (iHOT-33) in many patients.[51]

Generally, BMAC appears to be a safe treatment option. In the knee OA study (KL II-III), complications were modest at ~5.3%, including hematomas, numbness, swelling, and superficial infection; however, no major adverse events were reported.[47] In terms of the harvest site, studies harvesting from locations like the calcaneus show low complication rates, minimal/no pain at follow-up in most patients, no nerve injuries or wound complications.[52]

1.4 Retrospective Analysis on where BMAC Falls Short

Despite the reported results around BMAC there are still limitations that need to be overcome; currently the follow-up periods following BMAC treatments are fairly short ranging from a couple of weeks (3-12) to a few months (6-8) however, to both quantify and qualify the long term efficacy of the treatment follow-up periods closer to the 1-2 year mark would be more appropriate.

Furthermore, there is a wide range of variance in what is being referred to as ‘BMAC’ in these studies. Many studies report variable levels of volumes, different harvest sites (iliac crest, tibia, calcaneus, humerus, etc.), processing methods (no-spin, single, or double spins), injection protocols, etc. This makes group study comparison difficult at best and impossible at worst, the lack of standardization around the entirety of the BMAC therapeutic procedure direct hinders the long-term success of the treatment. Another issue is the reliance on patient reported outcomes, pain scores, or functional scores; specifically, there are a limited number of studies which perform image analysis, biomarker or histological assessment to meaningfully quantity rather than qualify the tissue repair properties of BMAC. In the same vein as biomarker and tissue analysis, there are even fewer studies which seek to directly assess how patient demographic factors such as age, sex, BMI, or other comorbidities such as smoking status are variably accounted for in the assessment of BMAC.

Chapter 2:

Framing the research gap and specific aims of this work

2.1 Specific Aims to Address Gaps in BMAC

To date, there remains a lack of comprehensive proteomic analyses performed on bone marrow aspirate concentrate (BMAC). Despite its growing use in clinical settings, both as a standalone treatment and in combination with other therapies, BMAC continues to be poorly characterized at the proteomic level. As a result, innate differences between patients, combined with the absence of standardized protocols for BMAC procurement, produce a therapeutic product with substantial variability. Furthermore, most clinical assessments of BMAC efficacy are qualitative rather than quantitative, which further limits our understanding of its biological mechanisms and therapeutic potential.

Accordingly, the goals of this project are twofold. First, we aim to characterize BMAC samples collected from distinct anatomical regions, specifically, the iliac crest and the humeral head, to determine whether the site of origin influences protein composition. By examining these proteomic profiles in the context of patient demographic factors such as age, sex, body mass index (BMI), and comorbidities (e.g., smoking), we seek to understand how these variables modulate the proteomic landscape of BMAC. Second, this work intends to establish a modular analytical framework that can serve as a scaffold for

future, larger-scale investigations into similar biologics and their clinical translation.

The central hypothesis guiding this study is that bone marrow derived from different anatomical regions possess distinct protein compositions, and that these profiles are further modulated by patient-specific demographic and lifestyle factors. To evaluate this hypothesis, three specific aims were developed:

- Aim 1: Determine the degree of separability between BMAC samples originating from either the iliac crest or humeral head and identify any key proteins which distinguish the BMAC samples from either site.
- Aim 2: Assess the influence of cigarette smoking on the BMAC proteome, both within marrow from the same region and across different anatomical sites by comparing current, former, and never-smokers. The goal is to identify protein biomarkers that may clarify whether BMAC remains an effective therapeutic in smoking populations.
- Aim 3: Identify sex-based differences in BMAC protein expression while accounting for confounding factors such as age and BMI, thereby determining whether biological sex contributes to proteomic variability within bone marrow aspirate concentrate.

Each of these aims will be addressed through a combination of inferential statistical analyses and machine learning–based approaches designed to uncover meaningful biological patterns within high-dimensional proteomic data.

2.2 Significance:

In the field of regenerative medicine, bio-injectable therapeutics such as BMAC have gained increasing prominence over the past several decades. Yet, BMAC remains something of a black box—a therapy widely used but not fully understood. There is a clear absence of studies that systematically examine the proteomic composition of BMAC and how it varies across patients and harvest sites.

This work seeks to fill that gap by conducting one of the most comprehensive BMAC proteomic analyses to date. To our knowledge, few studies have simultaneously examined BMAC from multiple anatomical regions while also integrating demographic and behavioral variables across a sizable dataset ($n = 123$ active samples, $m \approx 100$ unprocessed). Because BMAC can be generated from both the iliac crest and the humeral head, it is critical to understand the biological nuances that differentiate these sites and understand how such differences may influence therapeutic efficacy.

Beyond its biological insights, the significance of this work lies in its methodological innovation. By integrating proteomic data with statistical learning and computational modeling, this project bridges the gap between empirical observation and data-driven interpretation. In doing so, it establishes a framework through which future research can more effectively characterize therapeutics like BMAC.

Chapter 3

Exploration versus exploitation: Methods for examining BMAC

3.1 Overview of the methods used for proteomic analysis of BMAC

The therapeutic promise of BMAC described in the previous chapter relies on the interplay of cells, cytokines, and growth factors working in tandem to create what the literature would suggest is a potent regenerative environment. Despite its clinical success, the proteomic composition of BMAC is observed but not well characterized; specifically, researchers and clinicians understand what the key growth factors and anti-inflammatory molecules involved in BMAC are, there is a limited body of work which directly looks to characterize these proteins in the context of patient demographic factors, harvest sites, and comorbidities. To bridge this gap, a systemic understanding of both the proteins that mediate BMAC's therapeutic properties directly and tangentially is essential. Proteomic analysis therefore represents a critical next step in raising the bar with respect to BMAC as a therapeutic.

Proteomics can be broadly defined as the large-scale study of proteins with respect to their structure, function, and interactions in biological systems. With respect to BMAC, proteomic profiling allows researchers to look beyond just cell counts and clinical endpoints which only account for a portion of therapeutic outcomes. Understanding which proteins are enriched, depleted, or differentially

expressed provides insight into how patient individuality and harvesting techniques ultimately change the therapeutic potency of BMAC.

Historically, the number of proteomic studies involving BMAC is lacking; the main focus of BMAC centric studies often evaluate the cellular component of the therapeutic or the clinical outcomes involved with injecting BMAC. There are a variety of biochemical and immunological assays that are used in modern proteomic research including enzyme-linked immunosorbent assays (ELISAs) and Western blots, in addition to more advanced methods like liquid chromatography mass spectrometry (LSMS/MS). These techniques in particular have been used on BMAC in different capacities, which we will delve into in the next section.

3.2 Commonly Used Immunoassays

Conventional enzyme-linked immunosorbent assays (ELISAs) and multiplex immunoassays are some of the most common techniques. These methods are highly sensitive and specific for detecting known, pre-selected proteins. One such example comes from, Ragni et al. used an ELISA-based approach to screen 200 biomarkers (cytokines, chemokines, receptors, growth factors, and inflammatory molecules).[53] Likewise, Lan et al. applied a similar method using a 27-plex bead based ELISA to quantify cytokines and growth factors in BMAC.

Our own work (Herna et al.), sought to use antibody microarrays to evaluate 109 unique biomarkers in the BMAC proteome.[54] This technique shares the immunological nature of ELISAs but allows for diverse, high-throughput

screening of numerous analytes in parallel, which is more difficult to perform with more standard multiplex panels. Additionally, the antibody microarrays typically only consume a small volume of samples.

For targeted analysis of either a single protein or a small subset of proteins, Western blotting remains a valuable tool providing semi-quantitative data and confirmation of protein identity through its molecular weight. Hu et al. for example used the Western blotting technique to perform a semi-quantitative analysis on the presence of MMP-1, MMP-13, COL-I and COL-III as part of a study investigating BMACs ability to aid in prevention of hypertrophic scarring.[55] An alternative method with a steeper technical and analytical barrier to entry is liquid chromatography with tandem mass spectrometry (LC-MS/MS). As a cornerstone of unbiased, discovery-phase proteomics, MS can theoretically identify and quantify hundreds to thousands of proteins in BMAC plasma or cellular lysates without requiring pre-selected targets. Despite its power, it has been surprisingly underutilized in the analysis of human BMAC. A notable example comes from Ruoss et al., who performed a comprehensive single-cell transcriptional and proteomic analysis to build a comparative atlas of clinical-grade mesenchymal tissues, including BMAC.[56]

3.3 Theory Behind Antibody Microarrays

Antibody microarrays share the same immunological foundation as techniques such as ELISA or Western blotting, in that they rely on the specific binding between an antibody and its target antigen. However, the microarray

platform extends this principle to a high-throughput, multiplexed format capable of simultaneously screening and quantifying (some but not all) numerous proteins within a single sample.

In a typical, antibody microarray, capture antibodies specific to each analyte of interest are immobilized onto a solid substrate, such as a glass microscope slide, these form what are commonly referred to as spots or wells. Each spot functions as an independent reaction site designed to capture its corresponding target protein from the sample (Fig 3.1).

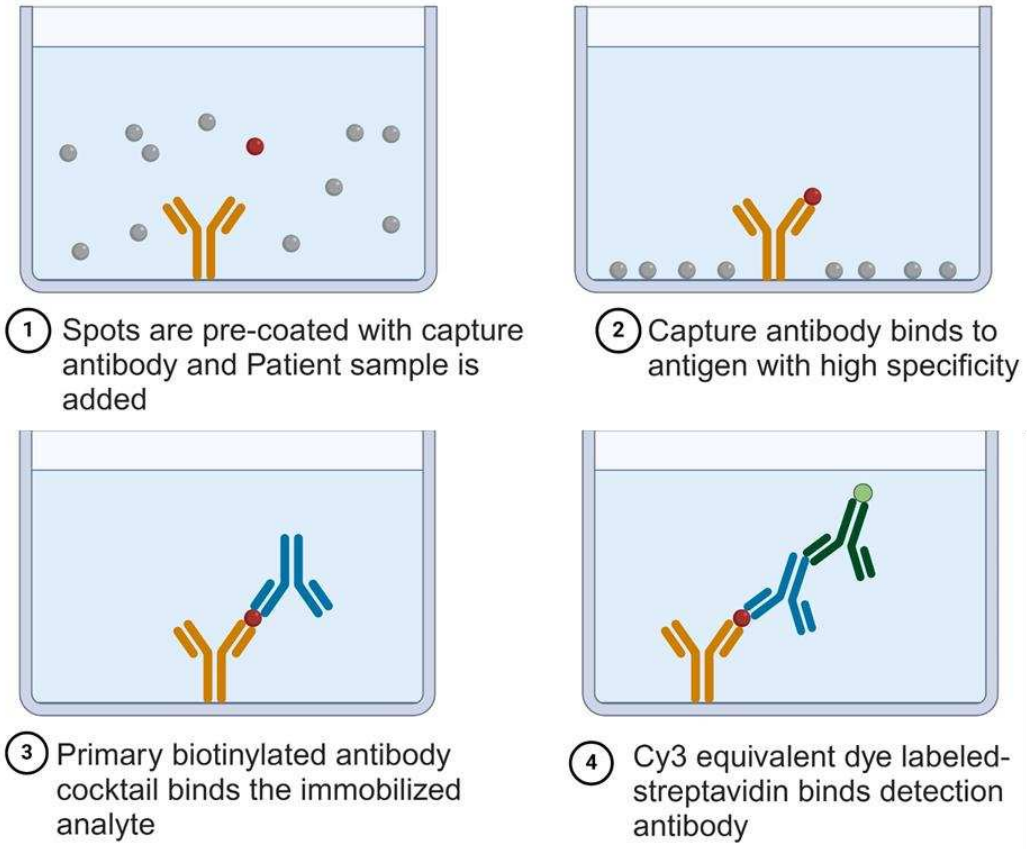

Figure 3.1 Overview of antibody microarray target capture: *Simplified schematic showing the theoretical capture mechanism for a single antibody microarray well. Created in BioRender. Jedlicka, S. (2025) https://BioRender.com/dinpkif.*

After the samples for analysis are applied to the spots, target proteins bind to their complementary capture antibodies. A primary detection antibody is then

introduced to bind the capture antigen, forming a sandwich-like complex. To amplify the detection signal, a secondary antibody or a fluorescently labeled streptavidin conjugate (like Cy3 or Cy5) are subsequently applied, which binds either directly to the primary antibody or to a biotin tag intermediary. The resulting fluorescent signal intensity is proportional to the amount of target protein bound at each spot. While this description applies to a single array feature, an individual microarray slide can contain both several individual arrays and hundreds of capture antibodies per array per slide; these screening are often performed in replicate for any given protein to ensure a certain level of robustness in the overall assay. This configuration enables the simultaneous profiling of dozens to hundreds of proteins using a relatively small volume of sample making antibody microarrays particularly powerful for proteomic analysis when the supply of a given sample is limited (Fig 3.2).

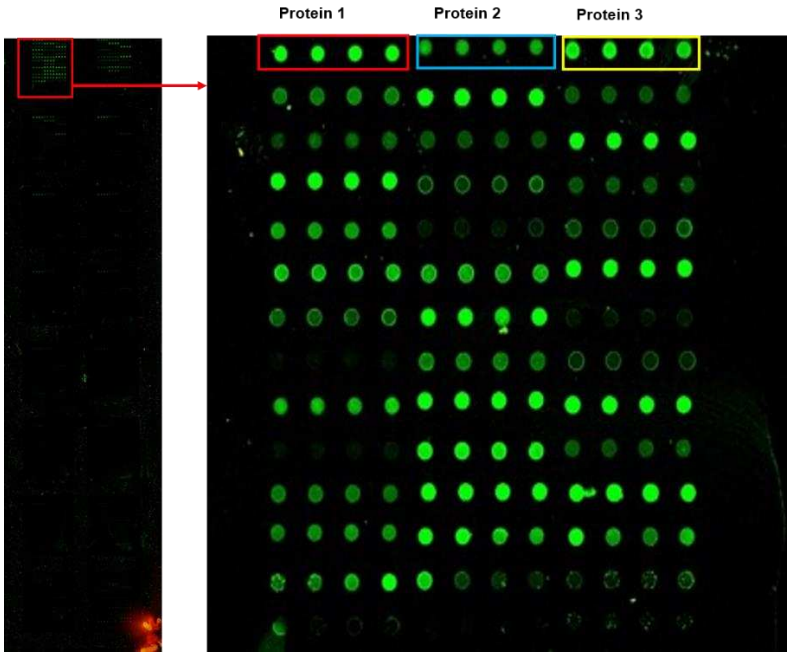

Figure 3.2 Antibody microarray slide with magnified view: *Image of a microarray slide containing 16 total microarrays and a zoomed in image of a singular microarray*.

3.4 Justification for the Use of Antibody Microarrays

The selection of the antibody microarray technique for this study was guided by both technical and practical considerations. Traditional immunoassays such as ELISAs and Western blots provide exceptional target specificity through antibody–antigen recognition but are inherently low-throughput. Screening 109 unique proteins in quadruplicate using these conventional methods would have required substantial time, sample volume, and laboratory resources. Conversely, discovery-phase platforms such as LC-MS/MS offer broad proteomic coverage but impose high technical barriers to entry, including the need for large sample inputs, sophisticated instrumentation, and advanced computational infrastructure. As outlined in Section 1.3, the total yield from a single BMAC extraction typically ranges from 3 to 8 mL, limiting the volume available for proteomic assays. Under such constraints, the antibody microarray platform offered an optimal balance, maintaining the immunological specificity of targeted assays while achieving the throughput and multiplexing capability of large-scale proteomic screens.

Another major advantage of antibody microarrays lies in their scalability. From an analytical standpoint, vertical scaling refers to expanding the total number of patients analyzed, whereas horizontal scaling refers to increasing the number of analytes screened per patient. The microarray format accommodates both dimensions efficiently. Standardized capture antibodies, replicate spots, and internal reference features create a stable assay environment that minimizes intra- and inter-chip variability. This design supports reproducible results across large sample sets, enabling robust statistical comparisons between patient

groups and across demographic or clinical covariates. Collectively, these features make antibody microarrays particularly well suited for comparative proteomic studies of BMAC, where the available sample volume is limited but the biological complexity demands high analytical depth

3.5 Literature Based Evidence in Support of Antibody Microarrays

Prior to adopting the antibody microarray platform, we evaluated its suitability based on precedent from both foundational and applied studies. Early, proof-of-concept work from the Department of Immunotechnology and CREATE Health at Lund University. These works, were led by authors Christer Wingren, Anders Carlsson, and Johan Ingvarsson, demonstrating the capability of antibody microarrays to generate high-dimensional proteomic profiles from clinically obtained serum samples. Over a series of investigations spanning multiple pathologies, including breast cancer [57], pancreatic adenocarcinomas [58], systemic lupus erythematosus (SLE) and systemic sclerosis (SSc) [59], and glioblastoma multiforme (GBM) [60], these researchers showed that antibody microarrays could discriminate between healthy and disease specific protein signatures using robust statistical learning approaches. Their work established a conceptual and methodological framework which could be extended to other biologically complex patient samples such as BMAC.

Complementary methodological groundwork was laid by Brian B. Haab and colleagues, whose 2001 publication validated a practical platoform for parallel detection and quantification of proteins in complex biological samples.

This study introduced several key technical innovations including two-color fluorescence labeling, printing antibodies on derivatized glass slides, and comparative assay designs [61], each of which remain a central component of modern microarray workflows. Haab's subsequent work in 2003 further demonstrated the translational relationship of antibody microarrays in cancer biomarker discovery, solidifying the platform's position between high-throughput discovery proteomics and clinically oriented immunoassays.[62]

Beyond oncology, antibody microarrays have been extensively applied to the study of autoimmune and inflammatory diseases, further supporting their adaptability to diverse biological contexts. For example, Quintana et al. screened serum samples from multiple sclerosis patients to link relevant biomarkers to different stages of disease progression[63], while Robinson et al. use custom antibody microarrays designed to mimic the protein profile of synovial proteins to characterize rheumatoid arthritis and systemic lupus erythematosus antibody responses as both an exploratory and methodological study.[64] These studies collectively demonstrate that antibody microarrays can capture disease-relevant proteomic and immunologic variation across diverse patient cohorts.

The unifying theme among these bodies of work is their shared goal of identifying biologically or clinically informative protein signatures from limited biological material. Together, they provide compelling evidence that antibody microarrays are an appropriate, efficient, and validated platform for high-throughput, targeted proteomic analysis. Within the context of our study, the limited volume of BMAC available per patient made this technique particularly

advantageous, as it allows for multiplexed detection across a wide dynamic range with minimal sample input.

However, implementing this technology also introduces computational challenges. The raw output of antibody microarrays is inherently high-dimensional, requiring systematic data preprocessing, normalization, and statistical modeling before meaningful biological interpretation can occur. As noted by prior investigators, effective analysis of such datasets depends on integrating principles from statistics, data science, and machine learning to uncover latent biological structure within correlated protein networks. These methodological considerations directly motivated the analytical framework described in the following chapter.

Chapter 4

Introduction to Data Science – Probability

4.1 A Framework for Proteomic Analysis:

The integration of data science into biomedical research has transformed how complex biological systems are studied and interpreted. In the context of proteomics, data science provides the analytical framework required to translate large-scale protein measurements into meaningful biological and clinical insights.[65] The datasets produced from antibody microarrays, for example, are high-dimensional, interdependent, and often exhibit subtle variations across patients, harvest sites, and processing conditions. Conventional univariate statistical methods, while useful for testing specific hypotheses, are insufficient to capture the multivariate interactions and latent patterns that characterize proteomics data.[66,67]

At its core, data science represents the convergence of statistics and probability, computer science, and domain-specific knowledge applied through the scientific method. Kelleher and Tierney illustrate this intersection with respect to the fundamental skills of a data scientist (Fig.4.1).[65]

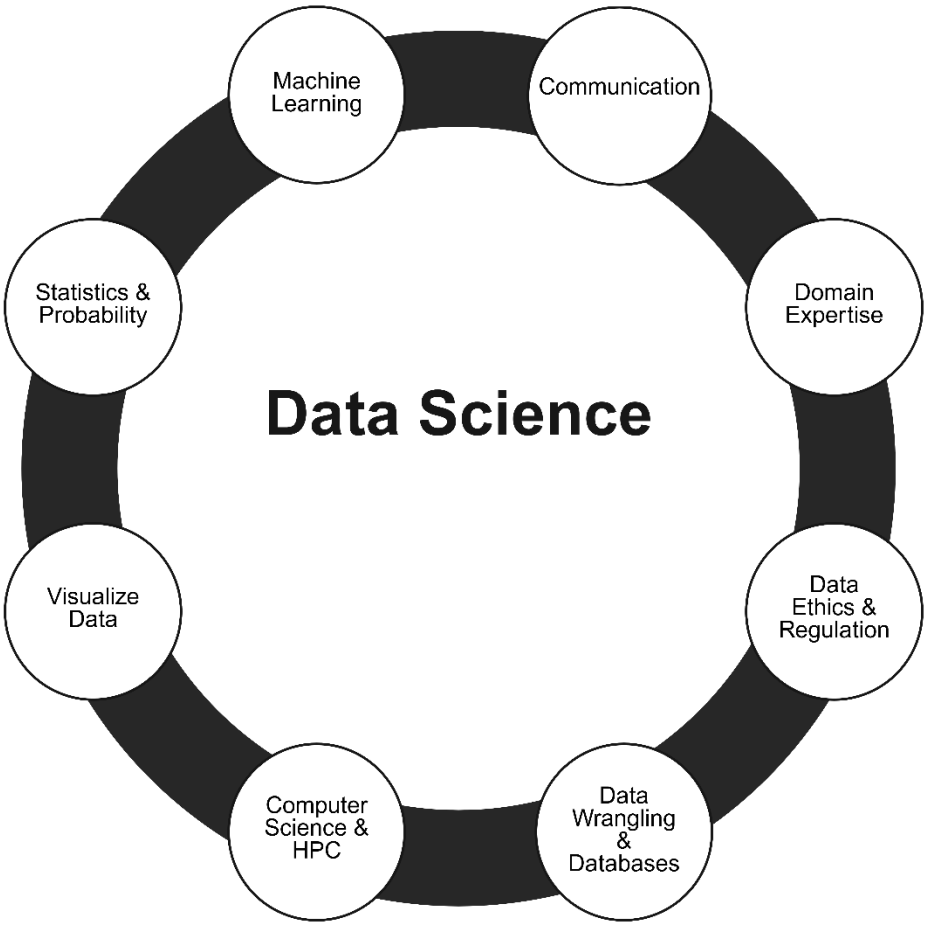

Figure 4.1 Data Science Skills: *Core skills required for a data scientist (adapted from Kelleher and Tierney, 2018).[65] BioRender. Jedlicka, S. (2025) https://BioRender.com/iszwe8h.*

Inside this framework, the goal is to simultaneously identify characteristics about individual components of the data, and then abstract those into relationships that broadly characterize the system as a whole.[66] For BMAC in particular, this means uncovering the network of proteins that collectively contribute to its regenerative and anti-inflammatory characteristics rather than viewing each as the result of singular analytes.

A standard data science workflow begins with data acquisition, preprocessing, and analysis.[65,68] In this study data acquisition corresponds to the semi-quantitative protein profiles generated from the antibody microarrays.[54] This

dataset captures 109 unique biomarkers per patient BMAC sample; these biomarkers span cytokines, chemokines, growth factors, and matrix regulators. Once collected, the data undergoes cleaning and normalization to ensure congruency and comparability across multiple workflows.[69] This process also helps to eliminate technical artifacts such as background noise, missing values, or measurement inconsistencies.

In our case, the preprocessing phase involves transforming raw fluorescence intensities into interpretable numerical representations, aligning replicates, and scaling values to reduce bias introduced by experimental variability.[70,71] Outlier detection and imputation are applied, when appropriate to, ensure that the respective presence or absence of a value does not have a misrepresented impact on the analysis as a whole. These steps seek to minimize as much of the technical and biological noise as possible without metaphorically squashing the variability that we are trying to observe. Through this process we hope to establish a foundation upon which high-level analyses can be performed.

Following the preprocessing phase, data validation ensures that the cleaned dataset adheres to the quality standards necessary for downstream statistical learning. In a biomedical context, validation often involves assessing data integrity, reproducibility, and compliance with analytical protocols. Generally, only after this step can the dataset be considered ready for exploratory or predictive modeling.

Finally, data analysis and modeling represent the stage at which hypotheses are tested, and new knowledge can be generated. This may include

descriptive statistics, correlation analysis, feature selection, clustering, and classification. For proteomic studies like this one, machine learning algorithms are particularly useful because they can identify multidimensional relationships and patterns which are usually too complex to capture with more traditional methods.[65,72,73] By applying these computations techniques, it becomes possible to distinguish protein expression signatures associated with patient demographic factors anatomic harvest sites, or other clinical variables.

To summarize, data science provides a structured and scalable framework for extracting biological meaning from complex proteomic datasets. When applied to BMAC, it enables the transition from empirical observations to a more quantitative perspective to better characterize therapeutic potency. This chapter subsequently details the analytical framework necessary to intellectually engage with the analysis of our work; an introduction to probability in the context of p-values and fundamentals of machine learning are laid out in an effort to guide the reader.

4.2 Expanding fundamental concepts of probability

Generally speaking, humans have an innate conceptual understanding of probability. If faced with a choice such as deciding to bring an umbrella with you on a dark and cloudy day, many would do so; if told to choose who we thought would win a race, or how well we may do on an exam, we could make reasonable guesses based on a combination of past data or experience and an awareness of the situation at hand. This probabilistic intuition is well described by

the classical work, *How to Improve Bayesian Reasoning Without Instruction: Frequency Formats*, by Gerd Gigerenzer and Ulrich Hoffrage.[74] Despite the innate intuition that we possess, gaps in fundamental knowledge of probability allow for exploitation, such as the famous Monty Hall problem.[75]

The Monty Hall problem is generally defined as: Suppose you are on a game show. There are three closed doors (1, 2, and 3), and behind each door is a car, a goat, or another goat. The host knows what is behind each door. You pick door 1. Before revealing the contents of door 1, the host reveals that there is a goat behind door 3 and offers you the option to switch your choice from door 1 to door 2. The question becomes, is it advantageous for you to swap your choice? Without delving too deep into the calculations, the short answer is yes, it is always worth swapping from door 1 to 2, given the choice. The intuitive answer is described as looking at the instances where the player (you) wins by staying with your original choice of door 1 and comparing that to the number of times you would win had you swapped doors. This case is outlined in the table below.[75]

Table 4.1 Monty Hall Problem: *Outlining the choices for the Monty Hall Problem*

Players Choice	Door with the prize	Door the host opens	Stay or switch
1	1	2 or 3	Stay
1	2	3	Switch
1	3	2	Switch

This question was the subject of quite a bit of controversy in the 1960s due to its subtly deceptive nature; However, it teaches us a valuable lesson in that our assumptions about the question greatly affect the way in which we go about solving the problem.[76]

With that example in mind, we will walk through the basics necessary to understand probability as it is the backbone of p-values that we observe in statistical hypothesis testing as well as the essence of machine learning. Probability is usually described as P of an event, E it is often expressed as:

$$P(E) = \cdots$$

The term event simply refers to an outcome such as flipping a coin, deciding if it will rain, choosing a flavor of ice cream, and so on.[77]

Consider the case where we have a bag of marbles containing blue, red, and green marbles. The probability of drawing any and all marbles is defined as the sample space, S. However, suppose we are only interested in drawing blue marbles, then we could look at the total probability of drawing any singular blue marble as a collection of outcomes. This collection of outcomes is typically referred to as an event, E; an event, E is a set of or collection outcomes rather than a single outcome.[78]

Take for example flipping two coins, we can define our sample space S as all of the possible outcomes from flipping our coins:

$$S = \{Heads - Heads, Heads - Tails, Tails - Heads, Tails - Tails\}$$

Once we have defined S, it is possible to then examine particular events, for example getting exactly one head, getting at least one head, first coin lands on heads, and coins are matching. Each of our events describes a different set of outcomes which can then be used to determine the probability of said events occurring. Take for example the one heads case; there are exactly two scenarios

in which we obtain a singular head. Therefore, we look at the possibilities divided by the total outcomes to obtain 50%

$$P(Exactly\ One\ Head) = \frac{\{Heads-Tails, Tails-Head\}}{\{\text{Heads}-\text{Heads}, \text{Heads}-\text{Tails}, \text{Tails}-\text{Heads}, \text{Tails}-\text{Tails}} = \frac{2}{4} = 0.5$$

Probabilistic calculations abide by 3 rules informally defined as:

1. $P(E) \geq 0\ ;\ Probabilities\ are\ positive$
2. $P(S) = 1\ ;\ Something\ must\ happen$
3. $P(A \cup B) = P(A) + P(B);\ When\ events\ cannot\ overlap, we\ add$

More formally, these rules are known as Kolmogorov's axioms of probability.[79]

We can convince ourselves by looking at a fair dice roll.[78] Let E_1 = Rolling X > 3 and let E_2 = Rolling 1. Based on our first rule, $P(E_1) = 3/6$ and $P(E_2) = 1/6$, they are not negative. The total possible rolls can be described as S = {1,2,3,4,5,6} or P(S) = 1. Finally, because we cannot roll a number greater than 3 and roll a one, we simply add the probability of each event occurring to get 2/3.[77]

The heuristic for solving these problems is as follows: Identify the sample space, define the events, apply rules. However, as the events become more complex, we must look for other cues to determine how we go about our calculations.
If we consider a poker game, some things we may ask ourselves are: "Did I get a king *or* an ace?", "Did I get a red *and* a face card?", "Did I *not* get a heart?". These events are no longer simple cases; they have evolved into combinations.

The "not" operation refers to the compliment (A^c) where the compliment to a probability can be defined as $P(A^c) = 1 - P(A)$.[78] Essentially the compliment is

encompassed by all cases which are not the case of interest. This relationship allows us to simplify what would ordinarily be complex calculations into rather simple ones.

The “and” operation looks at the intersection between events, denoted as $P(A \cap B)$.[78] Conceptually we can consider a Venn diagram and look at the overlapping region between circle A and B as the intersection between events and therefore what we wish to calculate.

The “or” operation is denoted as $P(A \cup B)$.[78] This operation is straightforward in that we are looking at P(A) + P(B), however, we must also account for the overlap between A and B in this calculation. Therefore, our calculation becomes $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

The “given that” or “conditional” operation is denoted as $P(A|B)$; This reads as the probability of A occurring given that B occurred.[78] We can define this more formally as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\ A\ and\ B)}{P(All\ B)}$$

The intuition behind this is that we take all the instances of events A and B overlapping and then divide out all the instances of B that could occur.

4.3 Bayes Theorem: A Quantifiable Way to Change Our Perspective

With a working framework of probabilistic fundamentals, we can now begin to look at deriving Bayes theorem. Consider two events, A and B. We write the probability of these events as an intersection,

$$P(A \cap B)$$

This intersection can be evaluated in terms of our conditional probability formula where,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \rightarrow P(A|B) * P(B) = P(A \cap B)$$

Now, if we reverse the order of events A and B occurring such that,

$$P(B \cap A)$$

We can again evaluate our expression as,

$$P(B|A) * P(A) = P(B \cap A)$$

We now have two expressions as a result of the intersection of two events being commutative,

$$P(A|B) * P(B) = P(A \cap B)$$

$$P(B|A) * P(A) = P(B \cap A) = P(A \cap B)$$

Since both expressions are equal to $P(A \cap B)$, we can set them equal to one another,

$$P(A|B) * P(B) = P(B|A) * P(A)$$

Finally, we simply divide the marginal probability of event B over to obtain Bayes Theorem,

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Defining each term we get:

- $P(A|B)$ represents the posterior probability, of the probability that event A happens given that B happened.
- $P(B|A)$ shows the inverse conditional probability (likelihood) that given that event A happened, what the probability of event B happening is.

- $P(A)$ represents the prior probability of event A happening, more intuitively this represents our initial belief that event A will occur.
- $P(B)$ represents the marginal probability that event B will occur.[80,81]

The practical use of Bayes Theorem is that it gives us a quantifiable way to take our current beliefs about an event happening, and updating those events based on new information. Bayes theorem is an explicit mathematical structure that tells us exactly how to update our beliefs about the world when encountering new evidence. Essentially, it enables us to move from a general, long-run frequency ("Event A occurs 20% of the time") to a specific, evidence-based belief ("Given what I know, event A should occur 20% of the time").[74,82]

4.4 What Probability Means in the Context of Data Analysis

Probability is the fundamental mechanism that allows us to quantify uncertainty and draw conclusions based on the information available to us. In data analysis, the interpretation is not monolithic; the way in which we define probability fundamentally changes the approach and interpretations in how we build our model of the world.[74,78]

Data is often treated as a realization of a random process. To model this, we use probability functions. For continuous data such as height or time, we approximate its distribution with a probability density function (PDF). For instances of discrete data such as counting, categories, or dice rolls we use a probability mass function (PMF).[77] These functions describe the likelihood of

different outcomes and allow us to calculate key properties like expected values (averages) or variance (deviations).[78]

A central challenge is that we must infer the true, underlying PDF from a finite sample of data. To make this problem tractable, statisticians employ a common approach known as parametric inferencing. Parametric inferencing involves assuming properties about the shape of the data or the distribution (e.g. Normal, Poisson, Binomial) and then using the collected sample data, estimate the parameters of the distribution such as mean and variance in the case of a normal distribution.[78,81]

The philosophical interpretation of probability is split into two schools of thought — those who believe probability is a frequency (Frequentists) and those who believe probability is a degree of belief (Bayesians) [81,82]:

Frequentist Probability: This framework interprets probability as the long-run frequency of an event. Here, frequentists believe that the population parameter (true mean for example) is a fixed and unknown value. Methods such as maximum-likelihood estimations (MLE) are often used to find an estimated parameter value which would make the observed data most probable.[83,84] This statistical inference typically feeds into hypothesis testing, often using p-values; the P-value represents the probability of observing data as extreme as, or more extreme than, the sampled data, assuming the null hypothesis is true. A small p-value ($p < 0.05$) suggests the data is unlikely under the given null hypothesis, meaning the null hypothesis is rejected. A common critique of this approach is

that it can tell us what is unlikely, but not directly the probability that a hypothesis is true.[85,86]

Bayesian Probability: This framework interprets probability as a subjective degree of belief. The parameter values are not fixed, but they are treated as random variables in themselves, each one containing their own probability distribution. Bayesian analysis as the name suggests, uses Bayes Theorem to update prior beliefs about a parameter with new evidence (the observed data) to form a posterior distribution. This posterior distribution provides a complete picture of our updated belief about the parameter, directly quantifying the probability that it exists within any given range. While this offers a more intuitive and direct form of inference, it requires specifying a prior distribution which can be an arduous task.

Despite the differences, both paradigms are powerful and widely used; often times the choice between one or the other is contextualized by the problem in question, availability of previous knowledge, and the desired conclusions.[81,82,85]

4.5 Real world example of Bayesian versus Frequentist

Through 2020, there was a global outbreak of the Corona virus which single handedly through the world into a global pandemic. In the race to break free from the pandemic status and return our previous normal, pharmaceutical companies raced to generate a modified vaccine. We will focus our attention on

two in particular, Pfizer's BNT162b2 mRNA vaccine and Moderna's mRNA-1273 SARS-Cov-2 vaccine.[87,88]

Pfizer's BNT162b2 mRNA using a Beta-Bayesian-Binomial Model

Pfizer conducted a Phase 3, placebo-controlled, observer blinded clinical trial with 43,448 participants over the age of 16. Participants were randomly assigned to receive two doses, 21 days apart, of either the BNT162b2 mRNA vaccine (n = 21,720) or a placebo (n = 21,728). After the second dose, only 8 COVID-19 cases were observed in the vaccine group, compared to 162 cases in the placebo group. This resulted in a vaccine efficacy of 95% with a reported 95% credible interval of 90.3% to 97.6%. Evidence of protection was observed soon after the first dose, with a vaccine efficacy of 52% before the second dose was administered. Critically, of 10 observed cases of severe Covid-19, 9 occurred in the placebo group, demonstrating high efficacy against the most serious outcomes.

The safety profile was characterized by short-term, mild-to-moderate reactogenicity, with a low and similar incidence of serious adverse events between the vaccine and placebo groups (0.6% vs. 0.5%). High efficacy was consistent across subgroups defined by age, sex, race, ethnicity, and underlying conditions.

Statistical analysis was conducted using a Beta-Bayesian-Binomial model. To maintain a type 1 error rate (false positive) of 2.5% despite interim analyses, the final posterior probability for vaccine efficacy exceeding 30% needed to be >

98.6%. The trial overwhelmingly surpassed this, showcasing a > 99.99% probability that the vaccine efficacy was > 30%. The 95% credible interval indicates that there is a 95% probability that the true efficacy lies between 90.3% and 97.6%.[87]

Moderna's mRNA 1273 SARS-Cov-2 Vaccine using a stratified Cox proportional hazard model

Moderna conducted a Phase 3, randomized, observer-blinded, placebo-controlled trial with 30,420 participants. Participants were randomly assigned to receive two doses, 28 days apart, of either the mRNA-1273 vaccine (n = 15,210) or a placebo (n = 15,210). In the final analysis, symptomatic Covid-19 was confirmed in 185 participants in the placebo group (56.5 per 1000 person-years; 95% confidence interval [CI], 48.7 to 65.3) and in 11 participants in the mRNA-1273 group (3.3 per 1000 person-years; 95% CI, 1.7 to 6.0). This resulted in a vaccine efficacy of 94.1% (95% CI, 89.3 to 96.8%; P<0.001). A key secondary endpoint was the prevention of severe Covid-19; all 30 severe cases, including one fatality, occurred in the placebo group, indicating 100% efficacy against severe disease.

Statistical analysis was conducted using a stratified Cox proportional hazards model. This frequentist-derived model assessed the mRNA-1273 vaccine efficacy by comparing the hazard (instantaneous risk) of developing Covid-19 between the vaccine and placebo groups over time. For the interim analysis, Moderna used the Lan-DeMets alpha-spending function with O'Brien–Fleming boundaries to determine the statistical significance thresholds for the

interim looks. In short, these methods are a standard way to conservatively adjust the p-value threshold required to claim vaccine efficacy for each analysis in an effort to control for type 1 error. The primary population analysis followed the per-protocol principle. Finally, the safety analysis—for which Moderna's clinical trial write-up cites a two-sided 95% exact confidence interval using the Clopper–Pearson method — reported an overall 94.1% effectiveness at preventing symptomatic Covid-19 and 100% efficacy at preventing severe disease, with a reported 95% confidence interval for efficacy ranging from 89.3% to 96.8%.[88]

4.6 Confidence Intervals Versus Credible Intervals

When examining the statistical approaches used by Pfizer and Moderna, it becomes evident that there is more than one way to evaluate uncertainty in vaccine efficacy. Both companies reported remarkably similar results, yet their analyses were grounded in distinct statistical philosophies.[87,88] The two key reported values from both Pfizer and Moderna lie in their 95% credible (Pfizer) and confidence (Moderna) intervals.

The 95% credible interval that Pfizer uses asks a simple question, "*Given the data we observed, what is the range of most plausible values for the true vaccine efficacy?".*[81,82] This metric therefore becomes intuitive by reading as, "There is a 95% probability that the true vaccine efficacy is between 90.3% and 97.6%". The 95% simply refers to a probability about our parameter value in this instance, vaccine efficacy. One of the primary strengths of using this approach is that in the posterior distribution of the event is informed by the prior distribution (initial

expectations about the vaccine efficacy). The use of this prior is often advantageous is it what is described as an “informed prior”, meaning that the value is informed by previous data, literature insight, expert opinions, etc.[81,89] The 95% confidence interval used by Moderna, however, asks a fundamentally different question; *“If we repeated our trial infinitely, what range of values would contain the true efficacy 95% of the time?*”.[81,90] This is often misunderstood because when a p-value is reported and a subsequent confidence interval follows it, the unindoctrinated take those items at face value [86]; However, the reality is that the reliability of the confidence interval is in the process or methods rather than the interval itself. With that, Moderna boasted a 95% confidence interval that the true vaccine efficacy was somewhere between 89.3% and 96.8%, meaning that over the long run or given infinitely many samples to their method, Moderna could say that they are 95% confident that their vaccine’s true efficacy existed within their interval.[88,90] The main limitation to this frequentist approach is the inability to inform your current stance based on a previous one i.e. there is not prior.

Despite the differences between one method or another it is important to note that both schools of thought are widely used in numerous domains to validate and learn about reported outcomes seen from observed data.[81,85] Contextually it is important to note that while the use of a prior can be essential in learning about a system, in the case of Pfizer versus Moderna, the prior was ultimately of little use relative to the scale of data that Pfizer had access to. As

such, Pfizer’s results were dominated by their posterior distribution, hence why both companies achieved very similar results.

Chapter 5

Introduction to Data Science – Statistical Learning

5.1 Fundamentals of Statistical and Machine Learning

As stated at the beginning of this chapter, data science is highly interdisciplinary, drawing from mathematics, computer science, probability theory, and so on. Machine learning, otherwise referred to as statistical learning, encompasses the same interdisciplinary nature. It can be stated that in the year 2025, data science is largely an umbrella term that encompasses skillsets like machine learning, however, it is important to note that they fundamentally rely on may of the same building blocks.

Statistical learning revolves around a stream of inputs and outputs. These inputs parameters are commonly referred to as predictor variables or features, and sometimes independent variables, often they are denoted as $X_1, X_2, \dots, X_i$. Outputs on the other hand are usually referred to as either response variables or dependent variables – these inputs and outputs generally will form pairs which are denoted as $(X_1, Y_1), (X_2, Y_2), \dots, (X_i, Y_i)$. Building on the probabilistic foundations established in chapter 4, statistical learning provides the framework for making data-driven predictions and interfaces.

We can broadly divide statistical learning methods into two buckets, supervised and unsupervised learning methods. These approaches differ with

respect to the presence of a response variable that is able to be mapped back to the predictor. Further considerations are made once a learning approach has been selected, the question we are trying to answer largely dictates the type of machine learning model that is employed. For example, questions that deal with quantitative data sometimes referred to as continuous data can be used as a regression-based model to look at how the input of one variable corresponds to the output of another. Alternatively, qualitative, or categorical data may be used to draw predictions about the behavior of a system. Both cases refer to a supervised learning approach where an input is mapped to an output. However, there are instances where we just apply inputs to a model with no outputs, this case refers to an unsupervised learning approach. The unsupervised learning approach is generally used for things like clustering data based on similarities to identify emergent properties of the data.[91,92]

5.2 Supervised Versus Unsupervised Learning Approaches

As introduced above, machine learning approaches can be divided into two paradigms, supervised and unsupervised learning. The critical distinction between these two comes from the presence or absence of a response variable, Y_i. In the supervised.

In supervised learning, the dataset acts as a teacher. For each input, X_i, we have a corresponding known output variable Y_i. The explicit goal is to learn a mapping function, f, from these inputs to the outputs such that we have, $Y = f(X)$. The core idea is that our model studies labeled samples to infer this relationship from training samples and is calibrated in a process called training or

fitting. Once fitted, our model is theoretically able to have novel data given to it, and in response output some type of predicted, $\hat{Y}$. Common examples can be seen below:

1. **Linear/Logistic Regression** – Predicting a continuous value like house prices or a binary category such as (spam/not spam).
2. **Decision Trees and Random Forests** – Classifying customers into groups or predicting equipment failures.
3. **Support Vector Machines (SVM)** – Identifying boundaries between classes such as in image recognition.
4. **Neural Networks** – Performing text-to-speech translations or vice versa.

Unsupervised Learning, on the other hand operates blindly, without a teacher. It is solely provided with the input data, X and no corresponding outputs. The goal is shifted from predictions to exploration, uncovering hidden patterns, structures, or intrinsic groupings within the data itself. Some common examples can be seen below:

1. **Clusters (K-Means, Hierarchical Clustering)** – Grouping similar data points together. For example, grouping customers based on their purchase patterns for targeted marketing or grouping genes involved in similar expression patterns.
2. **Dimensionality Reduction (Principal Component Analysis, t-SNE)** Compressing the data into a lower-dimensional space while attempting to preserve the variance within the data. These techniques have been a

staple in reducing complex high dimensional data into less noisy lower dimensional data prior to applying other models toward the output.

The choice between supervised and unsupervised learning is determined by the problem at hand and the data available.[93] However, a middle ground exists. In practice, data is often only partially labeled, giving rise to semi-supervised learning. This hybrid approach leverages a small amount of labeled data alongside a large volume of unlabeled data to develop a more robust and generalized model, often achieving higher accuracy than would be possible with the labeled data alone. Within supervised learning, the nature of the output variable further determines whether we approach the problem as regression or classification, which we will examine next.[91]

5.3 Deciding between Regression or Classification

The form of our response variable Y often guides our choices of what types of models or learning approaches to apply. For example:

Regression: Predicting house prices (continuous), forecasting stock prices, estimating patient recovery time

Classification: Detecting spam emails (spam/not spam), diagnosing diseases (healthy/sick), recognizing handwritten digits (0-9)

When Y is a quantitative variable, that is to say a numeric value such as age, height, income, or the value of a stock price; we often elect to use regression-based models. However, when Y is a class or a category we describe it as a qualitative variable; take for example marital status, purchases from a

particular brand of clothing, do you repay debt, or a medical diagnosis. These types of response variables usually use a classification-based model.

When we do opt for a classification-based model, our evaluation switches from looking at a raw numerical assessment of the model's predictions versus true outputs, and now we begin to quantify the fraction of how many instances the model was correct in classifying versus incorrect

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \hat{Y}_i)$$

where I now refers to an indicator variable that is equal to 1 when our prediction and true response variable to not match, and 0 when they do. This approach is referred to as the training error rate, or proportion of mistakes that are made when applying our model to the training data.[91,93]

Like many of our previous topics, the choice of model may not be so clear, and it is often guided by the questions being asked. A heuristic that most machine learning professionals abide by is to use a model that is both simple and interpretable whenever possible; a common pitfall is to always reach for the most complicated models right out of the gate.

5.4 Objectives of statistical learning

The previous sections outlined the decisions involved in model selection when initially looking at performing machine learning, however, they do not necessarily describe the centralized objective of the model itself.

The primary goal when performing any type of statistical learning is to approximate an unknown function which we will refer to as f, which links inputs

(X) and outputs (Y).[91,94] Generally, there are two main reasons we seek to estimate our unknown function f, prediction and inference:

Prediction: Generally, when X, inputs are readily available but the output, Y is not, we can look to predict Y using the following approximation,

$$\hat{Y} = \hat{f}(X)$$

the term $\hat{f}$ is often considered a black box because we generally do not care about the exact form of $\hat{f}$ assuming the approximation it provides is an adequate prediction for our true Y.

The accuracy of $\hat{Y}$ will generally depend on our error term, ϵ. Error can be broken into two groups, reducible error, and irreducible error. The name reducible error clues us into the fact that it is within our power to reduce this portion of the error term by optimizing various parameters of our model or selecting the most appropriate model itself. Alternatively, there is irreducible error, this can be conceptualized as the innate discrepancy between Y and $\hat{Y}$. Typically, irreducible error is the byproduct of variables in our system which cannot be accounted for and often go unmeasured for a variety of reasons.

James et al. describes the mathematical decomposition of the expected prediction error as follows[91]: Consider an estimate for $\hat{f}$ and set of arbitrary predictors X which yield the expression, $\hat{Y} = \hat{f}(X)$, then we assume these values are fixed such that the only discrepancy between Y and $\hat{Y}$ can only come from our error term, ϵ. We get the following expression:

$$E\left(Y - \hat{Y}\right)^2 = E\left[f(X) + \epsilon - \hat{f}(X)\right]^2 = \left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon)$$

Where the expected value of the squared difference between Y and $\hat{Y}$ is equal to the squared difference between the expected value of our true unknown function applied to our predictors plus our error term, minus the predicted function applied to the predictors. The latter expression can be reduced into two terms given my the right hand side of the expression; $\left[f(X) - \hat{f}(X)\right]^2$ describes the reducible error which can be reduced with the previously mentioned model or technical optimizations (This term can be further divided into bias and variance which will be covered later), $Var(\epsilon)$ on the other hand represents the calculated variance associated with our error term, otherwise known as our irreducible error.[91,93] It is important to note that the irreducible error will always dictate the upper boundary of potential model accuracy — more simply if the irreducible error is roughly 5% then our model can only ever maximally achieve 95%.

Inference: In this instance we want to know about the association between our inputs and outputs. This scenario does not treat $\hat{f}$ as a black box because we are interested in the exact form in this instance. Often we want to know how the inputs are associated with our outputs, because in practice, only a small subset of our predictor variables are truly insightful. Other times we may look to understand the relationship between our inputs and outputs with respect to their association; does a change in one input change the respective value of other inputs? The complexity of the relationship between inputs and outputs is also usually of interest, knowing if a system has a linear relationship (they rarely do in practice) can be useful for future endevours.[91]

Despite having these categories of prediction and inference, the true nature of real-world problems seek insights from both domains. Data sets and analysis can overlap based on the questions being asked with respect to the data. The previous prediction or inference frameworks offer a heuristic which statistical learning often abides by.

5.5 Parametric and Non-Parametric Modeling

For supervised learning approaches where we seek to map inputs and outputs, the way in which we do this involves dividing our data into two main units, training data and testing data. The training data is defined as the portion of our data set allocated for teaching our model how to estimate $f(X)$ with the goal that some function $\hat{f}(X)$ is approximately $Y = \hat{f}(X)$ for observations (X_i, Y_i). We can generally split the training modalities then facilitate our approximation, $\hat{f}$, as parametric and non-parametric:

Parametric methods are typically a two-step model approach where we first seek to find the form of f by assuming its form. After we select a model, then we must implement a protocol that uses the training data to fit the data we have to the model in question.[91]

For example, assuming a linear model,

$$f(X) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

because we assumed that our unknown function has a linear form, our job becomes simplified. Now rather than needing to approximate an arbitrary function with p-dimensions, we need only estimate the p+1 coefficients,

$\beta_o, \beta_1, \beta_2, \ldots, \beta_p$. These coefficients are referred to as parameters in this instance, a common method to approximate these parameters is to fit the training data to the model using the least squares approach; note however, this is simply one way of fitting, but there are numerous other methods available.

The pros of using a parametric function is that this approach greatly simplifies our search for an appropriate $\hat{f}$ by allowing us to choose a "good enough" fit and thereby move to estimating our parameters directly. The con of this approach though is because we rely on a parametric family to draw our model approximation through an assumed form of $\hat{f}$, generally forced to accept the discrepancies between the unknown function f and our model approximation $\hat{f}$ with some degree of irreducible error.

This irreducible error can be dealt with by choosing a flexible model, i.e. a model which does not have a rigid form such as a linear model. This would, however, require more parameters to be estimated with respect to our model to draw an appropriate approximation of our unknown function, furthermore, with this increased estimation we start to run the risk of overfitting our model to the training data specifically.

Non-Parametric methods, on the other hand, do not look to draw an explicit comparison between existing functions to map back to our unknown function. Instead, they seek to estimate f by approximating points as closely as possible to the data provided, without becoming too specific to the training data nor too complex in the way of the shape of our function approximation. The obvious benefit to this approach is that it can accommodate a much wider range

of unknown functions because the approach is not reliant on a guide function. Generally, this avoids the poorer potential fits that may be observed in a parametric approach. However, one of the major limitations to the non-parametric approach is that it often requires a significantly larger number of observations (samples) to adequately draw its approximation again because we are not using a guide function to model our unknown form.

5.6 Assessing the quality of a model's fit

To evaluate how well a statistical learning approach performs, we must measure how closely the model's predictions align with the true outcomes given by our data set. This comparison provides a quantitative measure for assessing a model's accuracy and generalizability.

With respect to regression models, a common metric used to evaluate model performance is the mean squared error (MSE), described by James et al.[91]. MSE measures average of the squared differences between the observed outcomes and the model's predicted values and is given by the following:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}(X_i)\right)^2$$

Here, Y_i represents the true observed response for the ith observation and $\hat{f}(X_i)$ denotes the predicted response obtained from our model. A small MSE indicates that the mode's predictions are, on average, close to the observed values, whereas a large MSE would suggest a greater discrepancy between the predicted and actual values.

It is important to note that the MSE is typically calculated separately for the training and testing components of the data. The training MSE reflects how well the model fits the data it was trained on given by:

$$MSE_{train} = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \left(Y_i - \hat{f}(X_i)\right)^2$$

This value often underestimates the true prediction error because the model has already seen the full scope of the data. In contrast the testing MSE is described by the same equation as the training MSE, except it uses testing observations, and more critically, it is computed based on previously unseen data, therefore providing a more accurate assessment of the model's true ability to generalize beyond the scope of its training data.

This idea of training and testing data can be intuitively thought of as the same relationship that a study guide has for an exam. The study guide represents our training data, the model is calibrated by the provided answers which help it learn. The exam serves as the testing data, it reveals how well we can handle novelty by generalizing learned information to new questions.

Ultimately, the goal of evaluating model fit is to balance good performance on the training data with a strong capacity for generalization to novel data. A model that performs exceptionally well on training data, but poorly on testing data is said to be overfit. If we take a moment to refer the study guide analogy, an overfit model is akin to a student who only is able to complete the questions in a study guide. Conversely, a model which performs poorly on both training and

testing data may be underfit meaning it fails to capture any essential structure of the data set.

By monitoring these metrics and their behaviors model complexity changes, we can iteratively refine our approach to achieve the most accurate and generalizable representation of the data. A note to the reader, the above describes a high-level analysis of accuracy, however, as we progress onward to chapter 6, a nuanced view of model metrics beyond accuracy will arise.

5.7 Tradeoff between Bias and Variance

Mathematically, the balance between accuracy and generalizability can be expressed by decomposing the expected test error as follows [93]:

$$E\left(Y - \hat{f}(X)\right)^2 = Var[\hat{f}(X)] + Bias[\hat{f}(X)]^2 + Var(\epsilon)$$

In this expression the total expected prediction error is partitioned into three terms: $Bias[\hat{f}(X)]^2$ represents the systematic error introduced by approximating $f(X)$ with a simplified model. For example, when modeling the relationship between height and weight, we might assume a simple linear relationship where taller individuals weigh more. However, this assumption neglects factors such as diet or body composition that affect weight independently of height, leading to bias in the model's predictions.

$Var[\hat{f}(X)]$ represents the sensitivity of the models' predictions to fluctuations in the training data. In other words, it quantifies how much the model's output would change if it were trained on a different dataset. Ideally, we seek to minimize this variance to ensure that the model's performance is stable and not overly dependent on specific training samples. On the other hand, $Var(\epsilon)$ represents

our irreducible error as previously described this is a catch all term for all things unmeasured or unaccounted for between the model and reality.

The bias-variance tradeoff describes the intrinsic tensions between a model's complexity and its ability to generalize. Typically, as a model's generalizability increases, the bias tends to decrease because the model can better calibrate itself to the training data. However, this often comes at the cost of higher variance, as the model becomes more sensitive to random fluctuations within the data. Conversely, a model that is too simple exhibits a high degree of bias, but low variance, consistently missing important relationships or properties of the data set. This tradeoff encapsulates the core challenge of statistical learning, that is to find the optimal balance between model complexity that minimizes total expected error. Graphically, this is often represented by a U-shaped curve where the test error first decreases with flexibility due to reduction in bias, but then sharply increases again due to a rise in variance.

In practice, techniques such as cross-validation, regularization, and model comparisons are used to identify the point at which this tradeoff achieves its minimum expected value. Striking this balance is critical for developing models that are both accurate and generalizable.[91,93]

5.8 Model Tradeoffs: Accuracy versus Interpretability

While bias-variance tradeoff quantifies how flexibility affects predictive accuracy, it is also important to consider how model flexibility influences interpretability.

Flexibility refers to the general complexity of the model insofar as the order of the terms or operations involved to finalize the outcome. Conceptually this can be thought of in terms of adaptability, where a flexible model is able to accommodate a wide range of data sets for various tasks as opposed to more rigid model which is generally stricter in the types of data sets it can fit.
With this idea of model complexity comes the tradeoff between a model's interpretability and flexibility. As a model becomes more flexible in how it able to handle different data sets, it also becomes much more difficult to capture the relationship between inputs and outputs, and in general visualize the model. Models such as linear regression or lasso are structured and rigid insofar as their flexibility goes, however, they are exceptionally easy to visualize and interpret. On the other hand, something like a deep learning model (neural networks) or some flavors of the support vector machine are exceptional at handling complex data but lack the ability directly and cleanly visualize the results. Our general guideline is that as a model gains flexibility, it loses interpretability at the extremes.[91,93]

5.9 Concluding Remarks

While the concepts presented in this chapter are not an exhaustive look of probability or machine learning, they establish the essential groundwork for more advanced analytical approaches. These principles form the basis for interpreting complex data and serve as the intellectual scaffolding for the analyses that follow. The ideas discussed regarding probability, statistical inference, and the

fundamentals of machine learning, extend beyond proteomics. They represent universal tools for reasoning about data across disciplines. However, in the context of this work, they provide the conceptual foundation necessary to analyze and interpret the high-dimensional proteomic datasets derived from BMAC. Having established this framework, the next chapter applies these principles directly to our experimental data. In doing so, we transition from theory to practice, demonstrating how probabilistic reasoning and machine learning methods can uncover meaningful biological patterns within the BMAC proteome and ultimately advance our understanding of its therapeutic potential.

Chapter 6:

Support Vector Machine Directed Proteomic Analysis of Bone Marrow Aspirate Concentrate

6.1 Abstract:

The aim of this study was to analyze the proteomic differences in bone marrow aspirate concentrate (BMAC) between extraction sites, specifically the iliac crest and humeral head. Using support vector machine (SVM) models, SVM(linear) and SVM(rbf), we identified six key proteins (MCSF, CD14, CD40, CD163, Ephrin-A4, and Matrilin-3) as essential for distinguishing between these sites. The results indicate that site-to-site proteomic differences are likely influenced by the surrounding tissues of each extraction site. Higher levels of chondrogenic and osteogenic proteins in the humeral head BMAC in addition to immune-modulatory proteins in iliac crest BMAC suggest site-specific proteomic profiles. This study provides insights that may assist clinicians in selecting sites-appropriate BMAC for therapeutic applications and inform future research into BMAC proteomics.

6.2 Introduction:

In sports and orthopedic medicine, biological therapeutics have become more prevalent for their ability to enhance existing treatments for musculoskeletal injuries and defects.[8,95–97] One of the common therapies is bone marrow aspirate concentrate (BMAC) which involves aspirating and concentrating down small volumes of bone marrow and injecting the product into the target site.[43,96,98,99]

BMAC is known to have high concentrations of anti-inflammatory and anabolic growth factors such as platelet-derived growth factor-B (PDGF-B), as well as bone morphogenic proteins 2 and 7 (BMP), interleukin-1 receptor antagonist (IL-1RA) which have been shown to inhibit the catabolic activity of IL-1, mesenchymal stem cell, and other progenitor cells.[43,100] Several studies have demonstrated the effectiveness of BMAC with respect to the need for revision surgeries[101–104]; one such example comes from a 2022 metadata analysis conducted at the Mayo Clinic which analyzed the Mariner data set from PearlDiver patient records repository and found that across a 114 patient group, when administered at the time of a rotator cuff repair (RCR), BMAC was associated with a significant decrease in revision surgeries.[96]

Bone marrow aspiration procedures are typically performed in an operating room, under anasthesia.[100] In many cases, bone marrow is aspirated from the iliac crest; this is largely due to its relative abundance of marrow, yield of progenitor cells, and association with preferred patient outcomes.[41,44,45,105] While the iliac crest is often the preferred location for aspirating bone marrow for BMAC

preparations, there may be patients for whom aspiration from a different location may be preferable.[41]

In the case of a full thickness rotator cuff tear, the humeral head becomes an accessible source of bone marrow which is local to the site of repair.[43,97,100] Typically, extracting marrow from the humeral head is challenging due to the musculature of the rotator cuff, in addition to the other surrounding tissue which ordinarily block the humeral head. However, when a full thickness rotator cuff tear occurs, a gap opens at the musculotendinous junction.[97,106] This gap can be used to both attach a surgical anchor for repair as well as aspirate marrow from the humeral head. With this, a surgeon may not need to add a second procedure to extract marrow from the iliac crest, rather they could perform the full surgery (repair and BMAC) in a single surgical site. Our work specifically examines BMAC samples originating from either the iliac crest or the humeral head with the aim of characterizing and comparing the proteomes of each marrow to determine the degree of similarity they possess. As BMAC can be produced from both sites, it is important to understand the site-dependent biological characteristics of the marrow. This could enable treatment decisions that are more patient and condition specific.

To our knowledge there are no pre-existing works which examine how the extraction site of BMAC changes its proteomic profile and how that affects its regenerative and therapeutic potential. In this study, the differences between the proteomes of the iliac crest-derived BMAC and the humeral head-derived BMAC

were characterized with a robust support vector machine (SVM) learning analysis performed to highlight the nuances of each sample.

The aim of this work was to use a SVM to determine the degree of separability or classifiability between the patient BMAC samples based on their proteomes. This information allowed us to analyze the biochemical comparability of the marrow's regenerative potential based on their proteomic profiles with an emphasis on the relative levels of anti-inflammatory, anabolic, and catabolic cytokines between the iliac crest and humeral head derived BMAC samples (Fig 6.1).

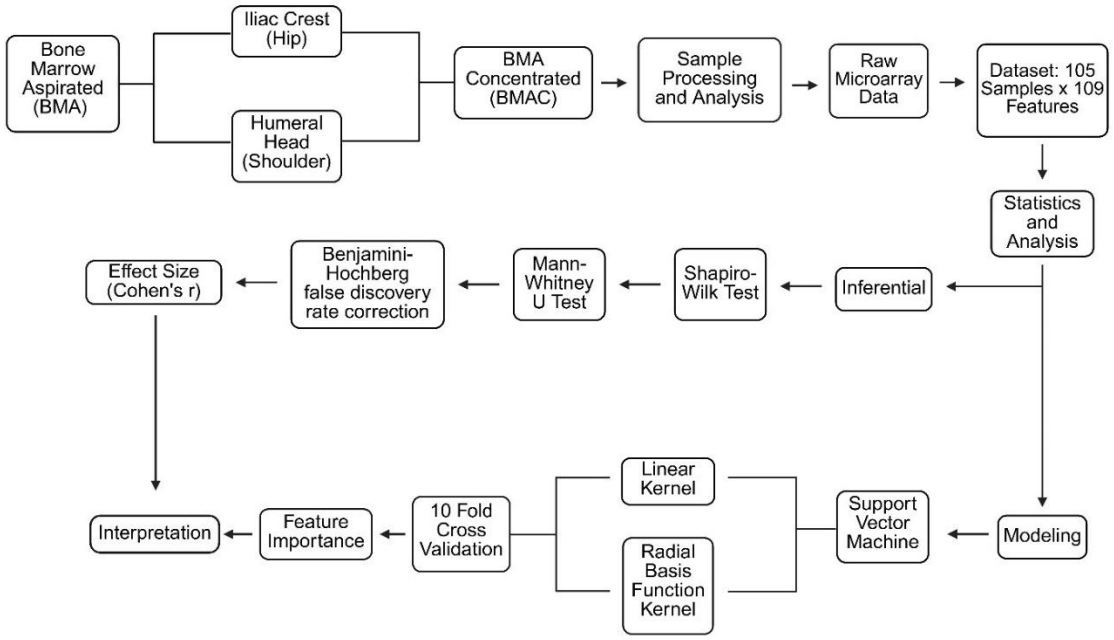

Figure 6.1 Schematic overview of the experimental and SVM analysis workflow used in this study: *Bone marrow aspirate (BMA) samples were collected from patients and processed into bone marrow aspirate concentrate (BMAC). Samples were then further separated into serum and cellular fractions, and the serum component was analyzed using antibody microarrays. Raw fluorescent data were processed in Python and Excel to build a proteomic dataset, which was then analyzed using inferential statistics and machine learning models to identify site-specific proteomic signatures. Created in BioRender. Jedlicka, S. (2025) https://BioRender.com/14yza7k.*

6.3 Methods:

Sample Acquisition: Bone marrow samples were aspirated (BMA) from patients (n=105) presenting with symptoms of joint pain (Table 6.1). BMA was processed into bone marrow aspirate concentrate (BMAC) by Sachdev Orthopedics as part of surgical intervention related to onset of patient pain using a commercially available centrifuge (Magellan, ISTO Biologics) to isolate the buffy coat containing the MSCs. The procedure was performed as a medical operation with informed consent. Leftover materials were deidentified and anonymized prior to being used as part of the subsequent analysis.

Ethical statement on samples: The specimens in question were collected under a routine clinical procedure, not as part of a clinical trial. The specimens were marked as waste, and further de-identified prior to receipt. As such, Lehigh University did not classify this as Human Subjects Research.

Table 6.1 Demographic Characteristics of Patients from Whom Bone Marrow Aspirate Concentrate (BMAC) Samples Were Acquired: *Demographic characteristics of patients from whom bone marrow aspirate concentrate (BMAC) samples were obtained. Data are presented for the iliac crest (n = 65) and humeral head (n = 40) cohorts, including sex distribution, age, BMI, and smoking status (current, former, never).*

Groups	Iliac Crest (65)	Humeral Head (40)
Sex, (M/F)	27/38	21/19
Average Age, (Years)	61.12 +/- 10.73	62.6 +/- 9.57
Smoking Status. (current, former, never)	10,13,42	6,11,22
Average BMI, ($m/kg^{2)}$	30.18 +/- 7.23	30.59 +/- 6.13

Sample Preparation: BMAC samples were separated into their serum and cellular components via centrifugation (1500g x 5 mins). Each BMAC serum sample was further processed using an Albumin & IgG Depletion SpinTrap (Cytiva) to remove excess solutes, immunoglobulins, and albumin proteins. Once stripped, each BMAC serum sample's total protein concentration was quantified using a bicinchoninic acid assay (BCA) using the following kit Micro BCA Protein Assay Kit (Thermo Scientific, 23227). All kits were used in accordance with their manufacturer protocols. After stripping and determining concentrations of each BMAC serum sample, samples were diluted at a final concentration of 100 ug/mL to standardize the concentration for microarray analysis.

Microarray slide preparation and scanning: BMAC serum samples at a final concentration to 100 ug/mL were loaded into four antibody microarrays (Table 6.2) from Ray Biotech; Human immune response array GS1 (GSH-IMR-1-2), Human inflammation response GS3 (GSH-INF-3-2), Custom G-series Select "Osteoclast" Array (GSH-CUST), Custom G-series Select "Cartilage" Array (GSH-CUST). The protocol for processing each set of arrays was provided by Ray Biotech.

Table 6.2 Map of Each Array Used to Characterize the Patient Bone Marrow Aspirate Concentrate Samples Based on 109 Unique Proteins: *Map of each array used to characterize the patient bone marrow aspirate concentrate samples based on 109 unique proteins. The maps classified as the "Human Inflammation Array GS3", "Human Immune Response Array GS1", "Custom G-series Select 'Osteoclast' Array", and the "Custom G-series Select 'Cartilage' Array".*

Array Protein Maps									
Human Inflammation Array GS3					**Human Immune Response Array GS1**				
BLC	Eotaxin	Eotaxin-2	G-CSF	GM-CSF	CD14	CD40	CD163	CRP	E-Selectin
I-309	ICAM-1	IFNg	IL-1a	IL-1b	Fas	FAS L	G-CSF	ICAM-1	IL-1a
IL-1ra	IL-2	IL-4	IL-5	IL-6	IL-1b	IL-2	IL-2 Ra	IL-4	IL-6
IL-6sR	IL-7	IL-8	IL-10	IL-11	IL-8	IL-10	IL-12p70	IL-13	IL-18
IL-12p40	IL-12p70	IL-13	IL-15	IL-16	Lipocalin-2	MCP-1	MCP-2	MIF	MIP-1a
IL-17	MCP-1	MCSF	MIG	MIP-1a	MIP-1b	OPN	PAI-I	PF4	Procalcitonin
MIP-1b	MIP-1d	PDGF-BB	RANTES	TIMP-1	RAGE	Resistin	ST2	Thrombomodulin	TNFa
TIMP-2	TNFa	TNFb	TNF RI	TNF RII	TREM-1	Troponin I	uPAR	VCAM-1	VEGF
Custom G-series Select "Osteoclast" Array					**Custom G-series Select "Cartilage" Array**				
BMP-4	CD109	EphA2	Ephrin-A4	FLRG	Aggrecan	bFGF	bIG-H3	BMP-2	BMP-5
Flt-3L	FSH	IFNg	IL-12 p40	IL-17	BMP-7	BMP-8	BMP-9	BMPR-IA	BMPR-IB
IL-20	IL-23	IL-23 R	ILT2	MCSF	BMPR-II	CHI3L1	FGF-4	FGF-6	FGF-9
M-CSF RA	MIP-1a	PTH	RANK	TGF-b1	IL-17F	Leptin	LRP-6	Lumican	Matrilin-3
TLR3	TLR4	TNFa	TRANCE	TREM-2	MBL	MMP-13A	NOV	sFRP-3	Thrombospondin-5

Data Pre-Processing: Once processed, each array was loaded into Genepix 4000B Microarray scanner (Molecular Devices) and imaged using the "GenePix Pro 6 Microarray Acquisition & Analysis Software". Proteins for each array type (Table 6.2) were analyzed in quadruplicate. The fluorescent signal intensity values from 'F532 Mean – B532' for a given protein were included if the value of the spot did not exceed 1.2x the median value of the collective spots for a protein.[107] The average number of replicates included for the Inflammation,

immune response, cartilage, and osteoclast arrays across the entire dataset were 3.5, 3.4, 3.5, 3.3 respectively (max is n=4). All spots which met the inclusion criteria were averaged and a log2 transform was applied. Final values were normalized (equation 1) to the average of two control samples (Human Serum Cellect, Fisher BioReagents, BP2657-100) run on opposite ends of each array. All data processing took place in Spyder IDE using Python 3.12 and Microsoft Excel.

[Equation 1] $$Normalized\ Samples = \frac{Averaged\ Raw\ Sample}{Averaged\ Raw\ Control}$$

Comparative Inferential statistical analysis: The BMAC dataset was stratified by anatomical source: Iliac crest (n = 65 samples) and humeral head (n = 40 samples). All 109 proteins were tested for normality using Shapiro-Wilk tests (α = 0.05). Due to non-normality in >70% of the proteins in both groups, non-parametric Mann-Whitney U (MWU) tests (two-tailed, α = 0.05) were employed for group comparisons. To control multiple comparisons (109 proteins = 109 comparisons), a Benjamini-Hochberg false discovery rate (FDR) correction was applied. Effect sizes were calculated using Cohen's r where positive (+) values indicate higher expression in the humeral head and negative (-) values the iliac crest.

Support vector machine learning analysis: The dataset was comprised of 105 patient samples defined by 109 features with a single binary classification target variable (Bone marrow source).[58,59,108,109] To perform this task a support vector machine (SVM) was imported “from sklearn.svm import SVC”, and two versions of this model were used: SVC(kernel = ‘linear’, C = 1, class_weight = ‘balanced’,

probability = True, gamma = ‘auto’) and SVC(kernel = ‘rbf’, C = 1, class_weight = ‘balanced’, gamma = 0.37). The parameters C and gamma were assigned based on the grid search for theoretical optima and live testing.[110] K-fold cross validation where K = 10, samples were shuffled, and the random state was fixed for reproducibility to address model overfitting.[58,59,108,109,111–113]

Model Performance Evaluation: Accuracy scores were used to evaluate the model’s performance at a baseline. However, each model’s performance was further evaluated with a combination of precision, recall, F1-scores, receiver operating characteristic curve with the area under the curve (ROC+AUC), as well as Mathew’s correlation coefficient. Each of these were calculated from a confusion matrix and verified using modules from the sklearn.metrics library.

Feature importance identification: To identify the significance of each feature (protein) with respect to the SVM(linear) we examined the model coefficients (Table 6.5. and Fig 6.4.). These model coefficients act as guides for the relationship of a given feature to a class output (humeral head or iliac crest). The sign of the value indicates the class that the feature likely belongs to, and the magnitude of the value ranks it amongst the other proteins.[114] Permutation feature importance scores were used to evaluate the SVM(rbf), similar to the SVM(linear), the magnitude of the score indicates how important the feature was to the model (Table 6.6 and Fig 6.5).

Identification of differentiating proteome features: After analyzing the feature importance of each SVM model, the overlap between what appeared to be the most important features in the context of BMAC sample classifiability and

identified their functions as it related to BMAC was compared and analyzed (Fig 6.6 and Table 6.7).

6.4 Results and Discussion

A total of 105 patient-derived bone marrow samples were characterized using four distinct antibody microarrays (Table 6.2): inflammation, immune response, osteoclast, and chondrogenic arrays. These arrays were selected for their diverse analyte coverage and overlap with proteins implicated in BMAC's therapeutic mechanisms. Selection prioritized proteins with established roles in BMAC function (e.g., inflammatory cytokines, osteogenic growth factors) while balancing practical constraints on sample volume and throughput. Although other arrays could have been informative in isolation, our selection captures many of the canonical pathways associated with BMAC. Notably, the design of this study allows for future horizontal expansion (additional proteins) and vertical expansion (additional patient observations).

Conventional methods for analyzing biological data typically involve comparative inferential statistics.[58,59,108] We elected to perform said analysis in conjunction with our SVM directed approach. Shapiro-Wilk tests revealed that only 24.8% (27/109) of the proteins in the iliac crest group and 25.7% (28/109) in the humeral head group followed a normal distribution, with just 7 proteins common between groups.[115,116] Given ~75% of the proteins followed a non-normal distribution, MWU tests identified 24 proteins with a raw p-value <0.05, and after FDR correction 7 proteins with an adjusted p-value < 0.05. The MWU test's raw p-value reflects a 5% probability that the observed differences between

the iliac crest and humeral head were by chance; with 109 proteins being tested, FDR correction was required to limit false discoveries. Cohen's r effect sizes contextualize the biological relevance where the magnitude of the value indicates stronger group separations, and the direction of that separation is indicated by the positive or negative signs aligned with humeral head and iliac crest respectively (Fig 6.2 and Table 6.3).[117]

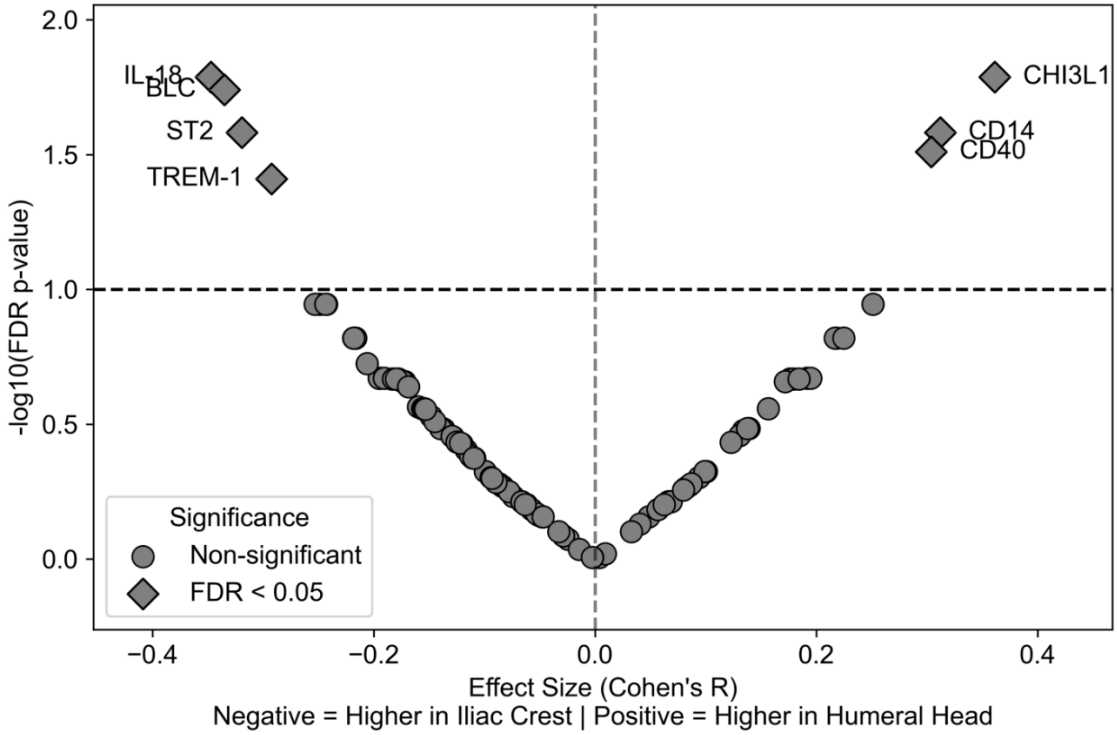

Figure 6.2 Volcano plot of differential protein expression between iliac crest and humeral head BMAC samples: *The y-axis shows the -log10 transformed p-vales after Benjamini-Hochberg false discovery rate (FDR) correction (n = 109 samples). Proteins above the significance threshold (y=1.0, FDR-adjusted p < 0.05) are statistically significant. The x-axis represents the Cohen's r effect sizes where Positive values: Higher expression in humeral head | Negative values: Higher expression in iliac crest. Black diamonds indicate FDR-Significant proteins (p <0.05), gray circles denote non-significant proteins.*

Table 6.3 Inferential Statistical Analysis of Patient Cohorts: *Differentially expressed proteins between iliac crest and humeral head BMAC samples. Raw P-value: Mann-Whitney U test results (unadjusted), FDR P-value: Benjamini-Hochberg corrected significance (adjusted), Cohen's r: Effect size magnitude/direction where Positive values: Higher expression in humeral head | Negative values: Higher expression in iliac crest. Bolded rows indicate proteins that remained statistically significant after FDR correction (p<0.05). Proteins sorted by raw p-value (most to least significant).*

Protein	Raw P-value	Cohen's r	FDR P-value
BMP-7	0.0463	-0.1945	0.2134
BMP-9	0.0471	-0.1938	0.2134
CHI3L1	**0.0002**	**0.3613**	**0.0163**
CD14	**0.0012**	**0.3120**	**0.0262**
CD40	**0.0017**	**0.3039**	**0.0309**
CD163	0.0250	0.2183	0.1515
CRP	0.0097	0.2511	0.1135
IL-4	0.0255	-0.2177	0.1515
IL-10	0.0264	-0.2164	0.1515
IL-12p70	0.0456	-0.1951	0.2134
IL-18	**0.0003**	**-0.3471**	**0.0163**
Procalcitonin	0.0105	-0.2486	0.1135
ST2	**0.0010**	**-0.3191**	**0.0262**
Thrombomodulin	0.0250	-0.2183	0.1515
TREM-1	**0.0025**	**-0.2924**	**0.0389**
BLC	**0.0005**	**-0.3349**	**0.0182**
Eotaxin	0.0092	-0.2531	0.1135
GM-CSF	0.0125	-0.2428	0.1135
TIMP-1	0.0123	-0.2434	0.1135
TNFb	0.0346	-0.2061	0.1886
CD109	0.0259	0.2170	0.1515
EphA2	0.0463	0.1948	0.2134
M-CSF RA	0.0210	0.2247	0.1515
TRANCE	0.0255	-0.2180	0.1515

While conventional inferential statistics (e.g. Mann-Whitney U tests) identify individual protein differences, they assume independence between features and struggle to model complex, high-dimensional relationships which are innately part of proteomic data. Given the moderate sample size of our dataset (n= 105 total samples) and the correlative nature of protein expression patterns, we employed an alternative support vector machine (SVM) analysis.

Comparison of the BMAC proteome in iliac crest and humeral head samples could be thought of as looking at the separability of the proteomes of both groups from one another. Ultimately, this becomes a classification task to identify similarity (or similar proteins) between each BMAC sample. The support vector machine (SVM) machine learning model was selected for its classification ability in conjunction with use cases in similar biological classification problems.[58,59,108,109,118]

Fundamentally, the SVM looks to plot a decision boundary between two classes of data which allows for classification from one class to another. The decision boundary is often referred to as a hyperplane and is oriented in such a way that maximizes the distance of the boundary from the closest data points of each class (support vectors). When a linear decision boundary (Fig 6.3A.) is insufficient in classifying the data (usually due to high dimensionality), a kernel function (Fig 6.3B.) can be implemented to expand the classification ability of the model by projecting the features into higher dimensional space.[91,109,112,119]

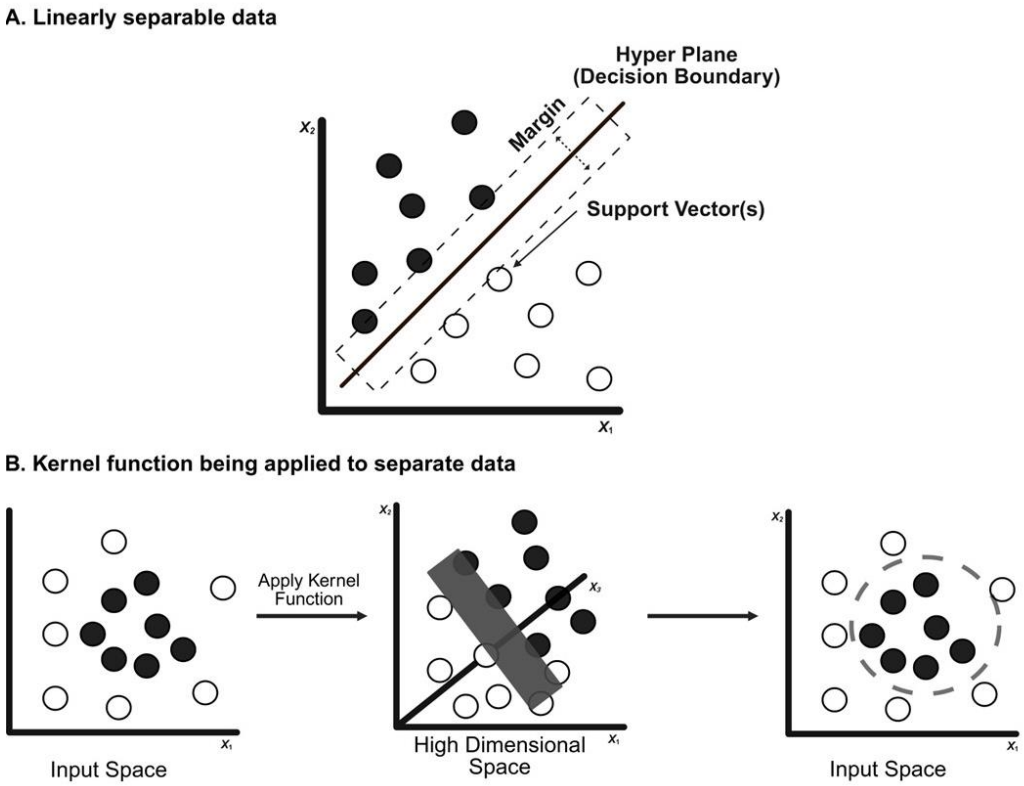

Figure 6.3. Schematic depicting the support vector machine classification decision boundary without the addition of kernel functions: *A.) Depicts the base case where the data being classified is linearly separable, B.) Illustrates how non-linearly separable data can be handled via a kernel function. Fig 6.3 Created in BioRender. Jedlicka, S. (2025) https://BioRender.com/0pogcqy.*

When fitting the model to a set of data the parameters “C” and “Gamma” are used to optimize the fit of the model with respect to the data being analyzed. C refers to the regularization parameter which controls the trade-off between the setting the distance between the decision boundary and the support vectors (maximizing the margin) and minimizing the classification errors. A low value for C allows for some misclassification in the base dataset, leading to better generalizability in novel datasets, higher values of C will prioritize correctness at the risk of overfitting. Typically values of C range from 1 to 10, where 10 is standard. Gamma is used to tune how much influence a single observation has over the SVM decision boundary. Lower gamma values result in smoother

decision boundaries that risk underfitting the data because the kernel has a wide decision radius. Conversely, a high gamma narrows the decision radius of the kernel meaning that the nearby points have the greatest effect. Baseline gamma values are typically decided by taking the inverse of the number of features in the dataset (1/109 in our case). It is important to note that when tuning gamma, it should be done with respect to C, high values for both C and gamma create overly complex decision boundaries which typically result in overfitting.[110]

Two variations of the SVM with different kernel functions (linear, rbf) were used. The average of a 10 fold cross validation accuracy assessment of each of these models yielded an average accuracy of 0.829 ± 0.111 for the SVM(linear) and 0.847 ± 0.122 for the SVM(rbf). This result indicates that the SVM models were able to predict the origin of the testing BMAC samples with room for generalizability to novel data.

The average confusion matrices were then extracted for each model. A confusion matrix is a 2x2 with four components true positive (TP), false positive (FP), false negative (FN), and true negative (TN) (Equation 2). These components are the backbone behind more comprehensive analytics such as precision/recall scores, F1 scores, or Mathew's correlation coefficients.[91,111] Take for example accuracy, this is a metric which can be broken down into Equation 3:

Equation 2: $$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \leftrightarrow \begin{bmatrix} True\ HH & False\ IC \\ False\ HH & True\ IC \end{bmatrix}$$

Equation 3: $$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$

Thus, one could evaluate each model's performance in a more comprehensive manner by looking at the TP, TN, FN, and FP components as seen in Equation 3, rather than solely looking at the accuracy score of a model.

Precision can therefore be used to look at the proportion of real true positive cases out of all positive (true and false) cases in the dataset (equation 4).[91,111] Under ideal circumstances precision will be 1, where the number of FP is 0. In the case of the models used here, the SVM(Linear) had a precision score 0.808 ± 0.173, whereas the SVM(RBF) scored 0.792 ± 0.164 meaning they were respectively able to identify marrow from the iliac crest ~80.8% and ~79.2% of the time.

Equation 4: $$Precision = \frac{TP}{TP+FP}$$

Recall (alternatively known as sensitivity) examines out of all of the perceived instances of a positive case, how many of those were real (Equation 5).[91,111] Again, an ideal value for this would be 1. The SVM was able to correctly identify marrow as originating from the iliac crest 0.825 ± 0.275 (Linear) and 0.825 ± 0.195 (RBF).

Equation 5: $$Recall = \frac{TP}{TP+FN}$$

Both the precision and recall can be used in tandem with one another to create an F1-score (Equation 6) which takes the harmonic mean (balances the scores) of the precision and recall scores and outputs a value ranging between 0 and 1. The closer the F1-score is to 1, the better the model is at maximizing positive cases and minimizing false cases, and therefore the better it is at binary

classification.[91,111] The SVM models produced an F1 score of 0.767 ± 0.180 (Linear) and 0.801 ± 0.164 (RBF).

Equation 6: $$F1\text{-}Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Another means of validating the model performance aside from the F1-score, is to plotting the rate of false positive cases versus the rate of true positive cases, to create a receiver operating characteristic curve (ROC). Once plotted, the ROC curve is typically evaluated by finding the area under the curve (AUC) where the values range from 0 to 1. A score of 0.5 is the equivalent of random guessing whereas a score of 1 is perfect classification; the SVM models scored 0.842 ± 0.118 (Linear) and 0.823 ± 0.167 (RBF).[91,111] However, the limitations of the ROC AUC are such that it does not provide the user with information about the positive predictive value nor the negative predictive value used by the classifier.[120] Adding the Mathew's correlation coefficient (MCC) (Equation 7) overcomes this shortfall. This maximizes the TP, TN,FP, and FN values, giving us a score between -1 and 1, where 1 corresponds to perfect prediction, 0 corresponds to a prediction made by random chance, and -1 corresponds to inversely perfect prediction, meaning all of the positive predictions were actually negative[120]. With our BMAC data, the SVM models scored 0.685 ± 0.124 (Linear) and 0.686 ± 0.261 (RBF). A summary of these results can be found in Table 6.4.

Equation 7: $$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP) * (TP+FN) * (TN+FP) * (TN+FN)}}$$

Table 6.4 SVM Performance Metrics: *Performance metrics for SVM models used to classify BMAC samples by anatomical source. Metrics include accuracy, precision, recall, F1-score, ROC AUC, and Matthew's correlation coefficient (MCC), with each presented as mean ± standard deviation over 10-fold cross-validation. Higher values indicate better classification performance.*

Metrics/Models	SVM(Linear)	SVM(RBF)
Accuracy	0.829 ± 0.111	0.847 ± 0.122
Precision	0.808 ± 0.173	0.792 ± 0.164
Recall	0.825 ± 0.275	0.825 ± 0.195
F1 Score	0.767 ± 0.180	0.801 ± 0.164
ROC+AUC	0.828 ± 0.124	0.842 ± 0.132
Matthews Correlation Coefficient (MCC)	0.685 ± 0.209	0.686 ± 0.261

When evaluating the performance of a machine learning model, overfitting is always a concern; this refers to a case where, the parameters of the model are adjusted (tuned) too specifically to training data, resulting in a loss of generalizability to novel data.[121] To combat the overfitting problem, we elected to perform a 10-Fold Cross Validation (K = 10) because with smaller datasets a larger K value helps achieve more stable and reliable model estimates.[121,122] The addition of K-fold cross validation (KCV) helps to reduce the model's dependence on a specific training-testing split for a classification prediction, because every sample is used in training and testing at some point, which serves to improve the robustness of the model by averaging the results of the K folds together for its final score.[122,123]

Additionally, the rationale for using a combination of SVM models was that the linear kernel has a low computational cost for training and testing the model, allowing for quick tuning and adjustments which could be further applied to the rbf kernel. In addition to its efficiency, the linear kernel has the added benefit of providing users with the feature model coefficients (Table 6.5. and Fig 6.4.) ultimately providing better interpretability of the model, which helped the model to

decide if a given sample is categorized as originating from the iliac crest or humeral head. Despite this, the linear kernel should be outperformed by a well-tuned rbf kernel as the features extend into higher dimensional space; however that trade in performance comes at the cost of not having access to the model coefficients.[113]

The model scores indicate a robust performance, demonstrating the linear model's ability to perform binary classification between BMAC samples from the humeral head and iliac crest. To better understand the interplay of the proteome in marrow classification, the individual proteins (rather than the model as a whole) were examined. To do so, the model weights were extracted from the SVM(linear) and ranked based on magnitude and sign. These model coefficients act in two ways: first, they show how a feature weight influences the model's prediction based on its sign. The SVM(Linear) model identified the top 10 ranking features most likely to be associated with the humeral head BMAC samples as ['CD14', 'CD40', 'CD163', 'IL-23 R', 'Lumican', 'Matrilin-3', 'CD109', 'IL-17', 'CHI3L1', 'M-CSF R'], and the top 10 features most likely to be associated with the iliac crest BMAC samples as ['BMP-5', 'Lipocalin-2', 'IL-18', 'IL-10', 'TIMP-1', 'MCSF', 'MIF', 'bFGF', 'Ephrin-A4', 'ST2'] (Table 6.5. and Fig 6.4.). These 20 proteins (features) provided insight into how exactly the model was making its predictions, and which features it deemed important with respect to classifying the marrows.

Table 6.5 Ranking Proteins From SVM(Linear): *Top 10 and bottom 10 protein features based on SVM (linear) model coefficients. Positive coefficients indicate association with humeral head samples, and negative coefficients with iliac crest samples. Coefficients were averaged over 10-fold cross-validation; standard deviations reflect variability across folds.*

Humeral Head Bias		Iliac Crest Bias	
Feature	**Model Weight ± Std Deviation**	**Feature**	**Model Weight ± Std Deviation**
CD14	1.309 ± 0.114	**BMP-5**	-0.664 ± 0.089
CD40	1.284 ± 0.134	**Lipocalin-2**	-0.669 ± 0.183
CD163	1.228 ± 0.094	**IL-18**	-0.685 ± 0.076
IL-23 R	0.834 ± 0.095	**IL-10**	-0.781 ± 0.126
Lumican	0.805 ± 0.085	**TIMP-1**	-1.106 ± 0.159
Matrilin-3	0.792 ± 0.111	**MCSF**	-1.212 ± 0.061
CD109	0.770 ± 0.098	**MIF**	-1.224 ± 0.133
IL-17	0.730 ± 0.142	**bFGF**	-1.238 ± 0.291
CHI3L1	0.669 ± 0.138	**Ephrin-A4**	-1.268 ± 0.130
M-CSF R	0.622 ± 0.090	**ST2**	-1.348 ± 0.197

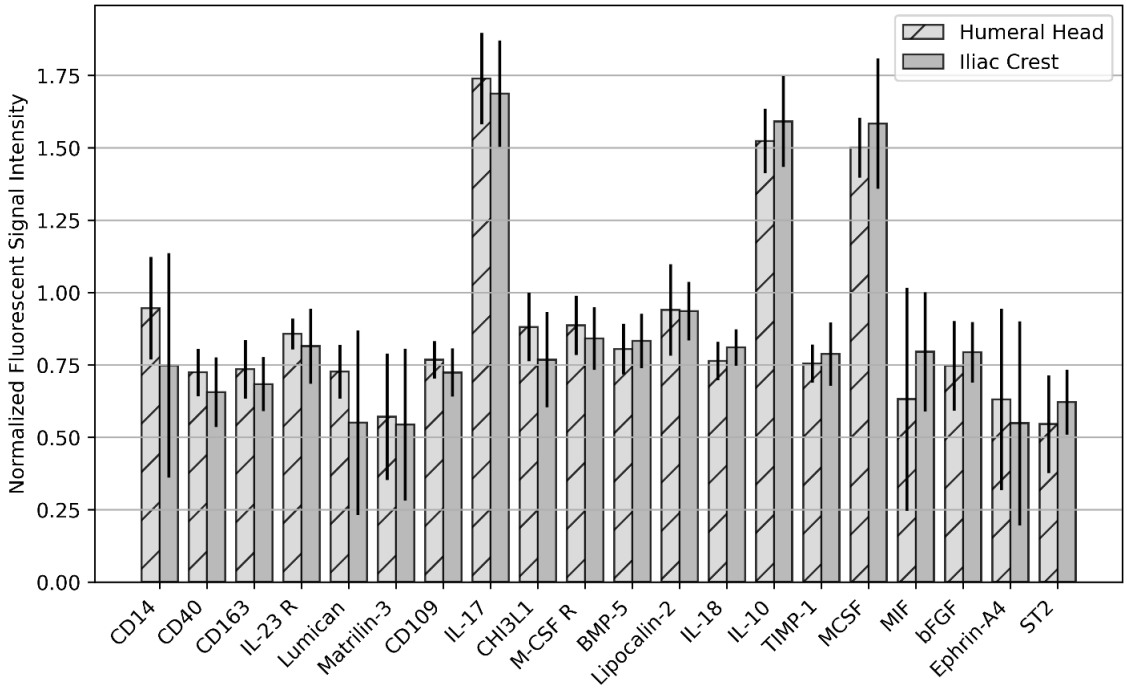

Figure 6.4 Protein signal Intensities from SVM(Linear): *Mean normalized fluorescent signal intensities with standard deviations of the 20 proteins that the SVM (linear) model coefficients suggested had the greatest impact on the model's classification of a BMAC sample being labeled as humeral head or iliac crest. The graph shows the levels that each protein expressed at in each BMAC group.*

Unlike the SVM (linear), the SVM (rbf) does not give us access to the model coefficients and the decision boundary cannot be cleanly visualized. Evaluating the SVM(rbf) is more challenging because what is gained in its enhanced classification ability, is lost in the models interpretability.[109,113,123,124] To visualize how the various features might be important to the SVM (rbf), the permutation feature importance (PFI) method was applied to each protein in our dataset; PFI randomly shuffles a single feature value, evaluates how this affects the model performance, and then assigns a score to the shuffled feature. This is then repeated across all the features in the dataset (Table 6.6. and Fig 6.5.). A high PFI score indicates that permutating a protein's value caused the model to perform worse, whereas a low PFI score indicates the model performed better. The intent of using this PFI method is to discern which features are important to the SVM(RBF) for making its predictions. It should be noted that the scores of a given feature do not speak to the degree of biological relevance that protein (feature) holds; rather the scores enable assessment of the values of the feature as it pertains to the classification task. For example, while IL-23 has a low permutation score relative to Fas-L, this does not mean that IL-23 is not important. Rather, this indicates that when the value associated with IL-23 was permuted, this perturbation did not significantly impact the model's ability to classify the BMAC samples. This distinction becomes important because the model is not speaking to the function of the protein in a system, but rather how alterations affect the biological fingerprint of the bodies they exist in. For this dataset, the top 20 PFI scores were for: ['FASL', 'Matrilin-3', 'Ephrin-A4', 'MIP-1a',

'BMP-4', 'MCSF', 'CD14', 'Resistin', 'IL-1b', 'il-12p40', 'IL-17F', 'PAI-I', 'sFRP-3', 'ICAM-1', 'IL-11', 'VEGF', 'CD163', 'IL-6', 'CD40', 'IL-23']

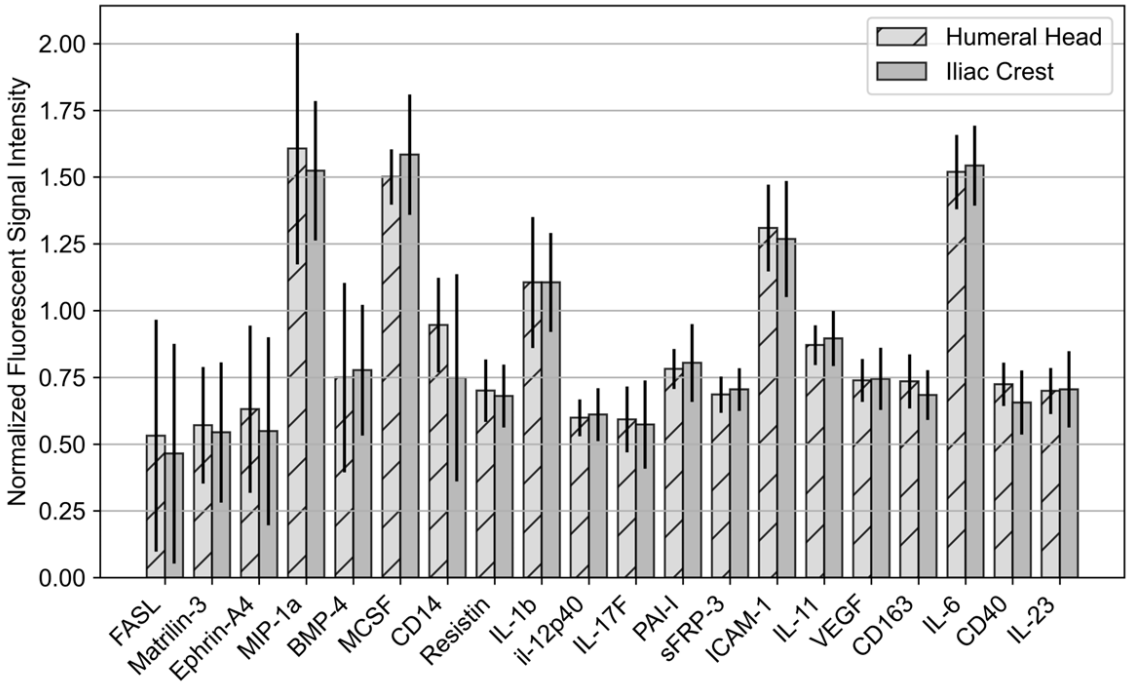

Figure 6.5* Protein signal Intensities from *SVM(RBF): *Mean normalized fluorescent signal intensities with standard deviations of the 20 proteins that the SVM (rbf) PFI scores suggested had the greatest impact on the model's classification of a BMAC sample being labeled as humeral head or iliac crest. The graph shows the levels that each protein expressed at in each BMAC group.*

Table 6.6 Permutation feature importance (PFI) scores for the SVM (RBF) model: *Scores reflect the change in model performance when a feature's values were randomly permuted. Higher positive values indicate that the feature had a stronger influence on model accuracy. Reported values are the mean ± standard deviation across 10 cross-validation folds.*

Feature	Permutation Score ± Std Deviation	Feature	Permutation Score ± Std Deviation
FAS L	0.025 ± 0.013	**IL-17F**	0.002 ± 0.002
Matrilin-3	0.013 ± 0.007	**PAI-I**	0.002 ± 0.002
Ephrin-A4	0.011 ± 0.008	**sFRP-3**	0.001 ± 0.001
MIP-1a	0.011 ± 0.011	**ICAM-1**	0.001 ± 0.006
BMP-4	0.007 ± 0.009	**IL-11**	0.001 ± 0.002
MCSF	0.004 ± 0.007	**VEGF**	0.001 ± 0.002
CD14	0.003 ± 0.012	**CD163**	0.001 ± 0.003
Resistin	0.003 ± 0.002	**IL-6**	0.001 ± 0.002
IL-1b	0.002 ± 0.004	**CD40**	0.001 ± 0.004
IL-12p40	0.002 ± 0.003	**IL-23**	0.001 ± 0.003

Of the 40 (out of 109) proteins examined between the SVM (linear) model and the SVM (rbf), there were 34 unique proteins identified as important for classification, with an overlap of 6 proteins between the models (Fig 6.6). Those 6 proteins being: Matrilin-3, Ephrin-A4, MCSF, CD163, CD40, and CD14.

Matrilin-3 is a well characterized extracellular matrix protein that helps in the formation of collagen networks.[125,126] Additionally, it plays a key role in cartilage development and ossification by helping modulate mesenchymal differentiation, and chondrocyte differentiation.[125,127–129] Matrilin-3's role in anabolic and catabolic activities appears to be contingent on its concentration; at concentrations of 100-200 ng/mL, it is shown to be able to modulate the activity of IL-1Ra, even if IL-1B is present in chondrocytes. Additionally, it has been shown to suppress MMP-13 activity in chondrocytes.[125,127] It is also able to inhibit BMP2 which is well known for its role in the hypertrophic differentiation of chondrocytes[125,130,131], however, it was shown when bound by Matrilin-3 this activity diminishes.[125,132] At higher concentrations of 5-50ug/mL Matrilin-3 increased the expression of MMP's, IL-1B, 6 and 8 in chondrocytes.[125,133] Further investigation into the expression of Matrilin-3 could prove useful in determining how much of an effect its expression has on the downstream targets.

Ephrin-A4 has well characterized roles in skeletal development as a negative regulator of osteoclast activity in addition to helping form the perichondrium.[134,135] Notably it acts as a synergistic regulator of IGF1 which has downstream effects on cartilage matrix degradation, and vascularization through upregulated expression of VEGF.[123,136,137]

MCSF has roles in the proliferation, differentiation, and survival of bone marrow progenitor cells.[138] This was found to be particularly true for osteoclasts which help to modulate bone mass; MCSF was found to be essential for the entire process of osteoclast differentiation spanning from the generation of precursor cells to the formation of mature osteoclasts.[139–142]

CD14 is a monocyte/macrophage differentiation antigen on the surface of myeloid lineage cells such as monocytes, macrophages, and dendritic cells.[143,144] Notably CD14+ monocytes play a central role in bone remodeling by way of differentiating into osteoclasts in the presence of MCSF and RANKL.[145] Monocytes will circulate out of the bone marrow in both inflammatory and homeostatic conditions and once in circulation can differentiate into osteoclasts.[146,147]

CD40 is a costimulatory signaling molecule in adaptive immunity which binds CD154 (CD40L) to provide secondary signals for T-cell dependent immune responses. Additionally, it has essential function in B-cell maturation and antibody class switching within the bone marrow.[148] In the context of inflammatory responses CD40 and its ligand increase the expression of cell adhesion molecules, pro-inflammatory cytokines, chemokines, and MMP's according to Ziegler et al 2019.[149]

CD163 is a well characterized scavenger receptor primarily expressed on macrophages; it modulates oxidative stress and inflammation by clearing hemoglobin-haptoglobin complexes and acts as a marker for M2-polarized macrophages.[150,151] With respect to bone marrow aspirate concentrate, CD163+

macrophages contribute to immunomodulation by promoting a regenerative microenvironment by secretion of anti-inflammatory cytokines such as IL-10 and pro-angiogenic factors like VEGF.[152,153]

Thus, to summarize, Matrilin-3 is involved in cartilage development and extracellular matrix modulations[125–128,130–132,154], Ephrin-A4, MCSF, and CD14 are involved in bone remodeling and osteoclast regulation[123,134,135,137–147]. CD40 and CD163 based on literature, are more involved in immunomodulation and inflammatory responses.[148–151]

We speculate that the site-to-site protein expression levels (Table 6.7) could be attributed to the tissue types surrounding the iliac crest and humeral head. The humeral head articulates in the glenohumeral joint which is a ball and socket joint, stabilized by the rotator cuff musculature and is lined with articular cartilage.[106,155] The iliac crest serves as an attachment site for muscles and connective tissue, hence it lacks a meaningful amount of cartilage.[156,157] Matrilin-3 had greater levels of expression in the humeral head BMAC samples, as did CD14, CD40, CD163, and Ephrin-A4. MCSF was slightly (~5.38%) more expressed in the iliac crest population. Given the general roles of each of these proteins and their relative expression across each BMAC sample; we believe that the tissues surrounding the bone marrow extraction sites ultimately influence the local proteomes of each marrow. Notably between the results of the Mann-Whitney U comparison test and the SVM models, two proteins CD14 and CD40 were mutually significant in each method.

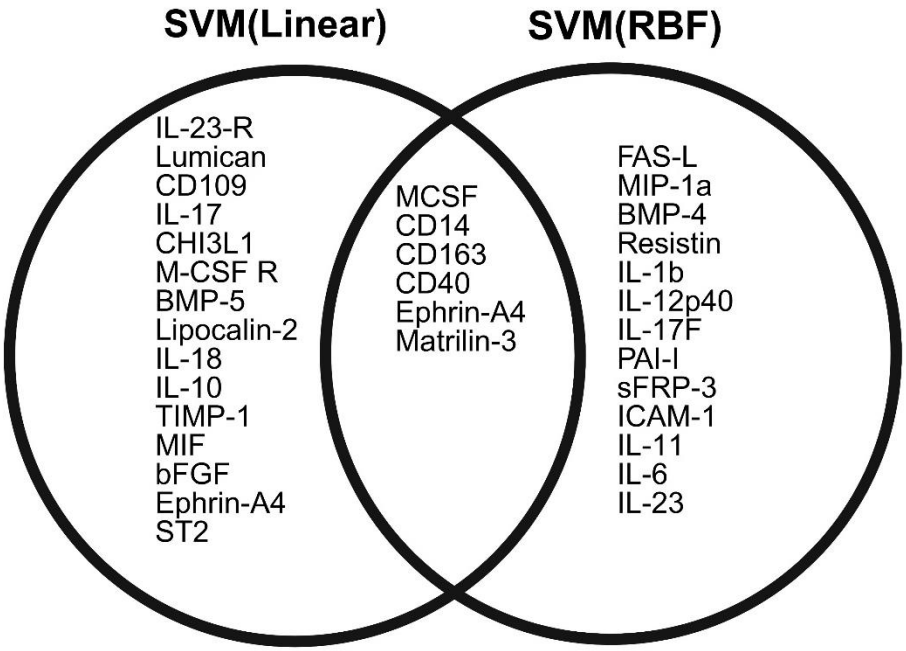

Figure 6.6 Venn diagram depicting the unique and overlapping top features between each SVM model: *The overlapping features were deemed significant enough to investigate further.*

Table 6.7 Comparison of average normalized fluorescent signal intensities for six proteins identified as important by both SVM models*: Percent differences reflect the relative expression between humeral head and iliac crest BMAC samples, highlighting site-specific proteomic trends.*

	Average Normalized Fluorescent Signal Intensity		
Protein	**Humeral Head**	**Iliac Crest**	**% Difference**
CD163	0.735	0.684	7.24
CD14	0.946	0.749	23.25
Matrilin-3	0.571	0.544	4.77
CD40	0.724	0.656	9.88
Ephrin-A4	0.631	0.549	13.99
MCSF	1.501	1.584	5.39

6.5 Conclusion:

Analysis of this dataset reveals that while the feature importance determined by the models do not speak to the biological relevance of said feature, they can provide us with the insight necessary, to look at the relationship between these proteins and bone marrow. This work was able to take BMAC samples from the humeral head and the iliac crest and identify ways in which

they differ by way of inferential statistics. These nuances between the BMAC's were further investigated with two flavors of SVM machine learning model to get a better grasp on which proteins in the dataset might be relevant. Based on varying feature performance metrics, we isolated 6 potentially key proteins which can be used to separate BMAC from the humeral head from that of the iliac crest. The results of this study can inform clinicians of the site-specific bone marrow nuances and help guide future proteomic studies.

To the best of our knowledge and current review of the literature, there are no studies which perform a direct proteomic analysis comparing BMAC samples derived from the iliac crest to the humeral head. Typically, when BMAC proteomic analyses are performed they revolve around a handful of key proteins (VEGF, BMPs, IL-1ra, IL-6, IL-8, TNF-a, TGF-B). Our study provides two advantages: First we screen a wide selection of proteins which have varying degrees of function within the niches of inflammatory pathways, immunomodulation, osteocyte and chondrocyte modulation. Second, our study provides a framework for others to apply a similar SVM directed approach to BMAC samples (derived from any location not just iliac crest or humeral head) and perform a deeper analysis by looking at the interactions between a myriad of proteins rather than single observations. In the next section, we expand upon this work by delving deeper into the interplay between patient demographic factors and relative protein expression with an emphasis on site to site differences.

Chapter 7:

Demographic Factors Affect Bone Marrow Aspirate Concentrate Proteome

7.1 Abstract:

Bone marrow aspirate concentrate (BMAC) is a promising orthobiologic therapy, yet the impact of patient demographics on its proteomic profile remains poorly understood. This study investigates how smoking status, age, sex, and BMI modulate BMAC's protein composition, focusing on 109 analytes across inflammation, immune response, and tissue regeneration pathways. BMAC samples from 120 patients (iliac crest: n = 77; humeral head: n = 43) were stratified by smoking status (current, former, never-smokers) and analyzed via antibody microarrays. Non-parametric statistical analyses (Kruskal-Wallis, Dunn's post hoc) with false discovery rate (FDR) corrections revealed significant proteomic alterations in iliac crest-derived BMAC from smokers, including upregulated pro-inflammatory cytokines (IL-17, TNF-α; $p < 0.05$, FDR-adjusted) and downregulated anabolic factors (TGF-β1, BMP-9). Former smokers exhibited intermediate profiles, suggesting partial recovery post-cessation. Sex-based differences (e.g., BMP-5, FGF-4) were noted but did not survive FDR correction. Humeral head samples showed minimal smoking-related changes, highlighting site-specific effects. These findings underscore smoking's catabolic influence on

BMAC's therapeutic potential and advocate for pre-treatment cessation protocols.

7.2 Introduction:

Bone marrow aspirate concentrate (BMAC) has emerged over the last 20 years as a unique treatment within orthobiologics. Through a combination of innately present mesenchymal stem cells, anti-inflammatory proteins, and growth factors, BMAC is able to deliver several powerful regenerative agents in a single injectable.[3,158]

To date the gold standard for BMAC acquisition is aspirating bone marrow from the iliac crest region of the hip and concentrating it down through one of many available centrifugation methods. To our knowledge the iliac crest has remained the gold standard due to its extractable volume of marrow, yield of progenitor cells (MSCs), and its association to preferred patient outcomes.[41,44,45,105]

Currently the role of MSCs (Mesenchymal Stem Cells) in bone marrow are well characterized, with works examining the osteogenic effects of bone marrow as early as the 19th century and later bodies of work demonstrating that osteogenic effects could be attributed to a small subpopulation of bone marrow cells.[24,159] MSCs are defined by the Mesenchymal and Tissue Stem Cell Committee of the International Society for Cellular Therapy as, "plastic adherent cells while being maintained under standard culture conditions, they must express CD105, CD73 and CD90, and lack expression of CD45, CD34, CD14 or

CD11b, CD79α or CD19 and HLA-DR surface molecules, as well as differentiate to osteoblasts, adipocytes, and chondroblasts *in vitro*".[28]

These cells host a variety of remarkable properties which when looked at as a whole, and become a sum greater than its parts; Intrinsic tropism more generally thought of as a homing mechanism is the method by which MSCs mobilize, adhere, and infiltrate into a given site of inflammation.[160] The anti-inflammatory activity of MSCs then allows for simultaneous upregulation of anti-inflammatory and downregulation of pro-inflammatory cytokines providing both local and systemic relief.[21,29] MSCs are also capable of apoptotic and immunomodulation through secreted cytokine regulation of B and T lymphocytes, natural killer cells, monocytes, neutrophils, and macrophages.[21,30,160] Aside from their regulatory effects, MSCs are capable of instigating tissue repair and regeneration by way of paracrine signaling; This results in significant regulation of cell proliferation, anti-oxidant activity, and differentiation.[31,32] Yet, despite our knowledge about the capacity of MSCs, there are still mixed results insofar as how patient demographic factors such as age, sex, BMI, smoking, and extraction site, affect the resulting concentration of these cells.[22,41,42,161]

The concentration of MSCs in BMAC are known to be affected by patient demographic factors, however, with respect to the BMAC proteome the effect of demographic factors is less characterized. The cellular component of BMAC is better characterized than its proteomic counterpart; Extensive proteomic analyses on BMAC are rarely performed despite many studies directly naming proteins such as VEGF, PDGFs, BMPs, IL-1ra, and TGF-B1 as some of the key

anti-inflammatory and growth factors housed in BMAC.[3,44,149] However, when BMAC proteomic analyses are performed they are typically done through the lenses of the MSC secretome, rather than the native proteins in BMACs “serum” component.[4,162–164] The general lack of standardization around the BMAC processing procedure, further propagates the issue of not truly understanding how patient demographic factors modulate both MSC and serum protein concentration.[44,165]

Our aim for this work is to evaluate the effects of smoking with respect to patient demographic factors (age, sex, bmi) to better understand how specific proteins may be differentially affected by a patient’s smoking status. There are numerous ways to go about parsing this data, but, given the catabolic nature of smoking in contract to the anabolic nature of BMAC we have elected to focus on how age, sex, and BMI, interact with smoking.

Consumption of cigarettes often results in many diseases including cancer and cardiovascular and lung diseases. According to the CDC as of 2022 roughly 1 in 5 adults (49.2 million) use cigarettes. Tobacco products remain as one of the leading causes of preventable disease and death in the United States.[166] Cigarette smoke itself contains reactive oxygen species (ROS) and a myriad cytotoxic compounds.[170–172] which have been shown to accelerate aging across a variety of tissues in the body.[167–169]

At the cellular level cigarette smoke exposure (CSE) has been shown to delay chondrogenesis in mice and inhibit osteogenesis in rabbits.[173,174] Furthermore, in MSC’s derived from adipose tissue, CSE exposure impaired the

migration and viability, altered the secreted protein profile, and had genetic variations throughout differentiation into osteocyte and chondrocyte lineages.[175] *In vitro* bone marrow MSCs were isolated from guinea pigs and exposed to cigarette smoke; authors found that short term CSE results in diminished MSC homing, proliferation, differentiation, and migratory abilities to be diminished.[176]

The damage that smoking causes can be summarized as catabolic in nature, corroding tissues to point of cellular senesce, apoptosis, or proliferative malfunction. However, the link between smoking and BMAC as a treatment is not explicitly well documented. Patients are advised to refrain from smoking and alcohol prior to an appointment for BMAC treatment. Currently what we understand is that smoking and BMAC have antagonistic effects in that smoking produces a catabolic response in the body, whereas BMAC produces an anabolic response in the body. For this reason, actively smoking patients are believed to receive reduced to little effectiveness from the BMAC therapy.

7.3 Methods:

Samples Acquisition: Bone marrow samples were aspirated (BMA) from patients (n=120) presenting symptoms of joint pain (Table 7.1). BMA was processed into bone marrow aspirate concentrate (BMAC) by Sachdev Orthopedics as part of surgical intervention related to onset of patient pain using a commercially available centrifuge (Magellan, ISTO Biologics) to isolate the buffy coat containing the MSCs. The procedure was performed as a medical

operation with informed consent. Leftover materials were deidentified and anonymized prior to being used as part of the subsequent analysis.

Ethical statement on samples: The specimens in question were collected under a routine clinical procedure, not as part of a clinical trial. The specimens were marked as waste, and further de-identified prior to receipt. As such, Lehigh University did not classify this as Human Subjects Research.

Table 7.1 Demographic Characteristics of Patients from Whom Bone Marrow Aspirate Concentrate (BMAC) Samples Were Acquired: *Data are presented for the iliac crest (n = 77) and humeral head (n = 43) cohorts and further stratified by patient smoking status (smoker, former smoker, or never smoker). The average age, sex, and BMI are reported for each smoking group.*

Bone Marrow Source	Smoking Status	Average Age (Years)	Sex (M | F)	Average BMI (KG/M^2)
Iliac Crest (77)	Smoker (15)	56.20 $\pm$ 8.03	7 | 8	28.72 $\pm$ 3.49
	Former Smoker (20)	63.20 $\pm$ 8.31	10 | 10	31.56 $\pm$ 5.16
	Never Smoker (42)	63.61 $\pm$ 10.17	14 | 28	30.84 $\pm$ 7.48
Humeral Head (43)	Smoker (9)	58.33 $\pm$ 9.86	8 | 1	29.99 $\pm$ 7.41
	Former Smoker (11)	64.45 $\pm$ 9.98	6 | 5	30.02 $\pm$ 5.47
	Never Smoker (23)	63.61 $\pm$ 10.17	10 | 13	30.77 $\pm$ 6.59
All (120)	Smoker (24)	57.00 $\pm$ 8.61	15 | 9	29.18 $\pm$ 5.12
	Former Smoker (31)	63.65 $\pm$ 8.79	16 | 15	30.94 $\pm$ 5.24
	Never Smoker (65)	62.85 $\pm$ 11.14	24 | 41	30.81 $\pm$ 7.10

Sample Preparation: BMAC samples were separated into their serum and cellular components via centrifugation (1500g x 5 mins). Each BMAC serum sample was further processed using an Albumin & IgG Depletion SpinTrap™

(Cytiva) to remove excess solutes, immunoglobulins, and albumin proteins. Once stripped, each BMAC serum sample's total protein concentration was quantified using a bicinchoninic acid assay (BCA) using the following kit Micro BCA™ Protein Assay Kit (Thermo Scientific, 23227). All kits were used in accordance with their manufacturer protocols. After stripping and determining concentrations of each BMAC serum sample, samples were diluted at a final concentration of 100 ug/mL to standardize the concentration for microarray analysis.

Microarray slide preparation and scanning: BMAC serum samples at a final concentration to 100 ug/mL were loaded into four antibody microarrays (Table 7.2) from Ray Biotech; Human immune response array GS1 (GSH-IMR-1-2), Human inflammation response GS3 (GSH-INF-3-2), Custom G-series Select "Osteoclast" Array (GSH-CUST), Custom G-series Select "Cartilage" Array (GSH-CUST). The protocol for processing each set of arrays was provided by Ray Biotech.

Table 7.2 Map of Each Array Used to Characterize the Patient Bone Marrow Aspirate Concentrate Samples Based on 109 Unique Proteins: *The maps are classified as the "Human Inflammation Array GS3", "Human Immune Response Array GS1", "Custom G-series Select 'Osteoclast' Array", and the "Custom G-series Select 'Cartilage' Array".*

Array Protein Maps									
Human Inflammation Array GS3					**Human Immune Response Array GS1**				
BLC	Eotaxin	Eotaxin-2	G-CSF	GM-CSF	CD14	CD40	CD163	CRP	E-Selectin
I-309	ICAM-1	IFNg	IL-1a	IL-1b	Fas	FAS L	G-CSF	ICAM-1	IL-1a
IL-1ra	IL-2	IL-4	IL-5	IL-6	IL-1b	IL-2	IL-2 Ra	IL-4	IL-6
IL-6sR	IL-7	IL-8	IL-10	IL-11	IL-8	IL-10	IL-12p70	IL-13	IL-18
IL-12p40	IL-12p70	IL-13	IL-15	IL-16	Lipocalin-2	MCP-1	MCP-2	MIF	MIP-1a
IL-17	MCP-1	MCSF	MIG	MIP-1a	MIP-1b	OPN	PAI-I	PF4	Procalcitonin
MIP-1b	MIP-1d	PDGF-BB	RANTES	TIMP-1	RAGE	Resistin	ST2	Thrombomodulin	TNFa
TIMP-2	TNFa	TNFb	TNF RI	TNF RII	TREM-1	Troponin I	uPAR	VCAM-1	VEGF
Custom G-series Select "Osteoclast" Array					**Custom G-series Select "Cartilage" Array**				
BMP-4	CD109	EphA2	Ephrin-A4	FLRG	Aggrecan	bFGF	bIG-H3	BMP-2	BMP-5
Flt-3L	FSH	IFNg	IL-12 p40	IL-17	BMP-7	BMP-8	BMP-9	BMPR-IA	BMPR-IB
IL-20	IL-23	IL-23 R	ILT2	MCSF	BMPR-II	CHI3L1	FGF-4	FGF-6	FGF-9
M-CSF RA	MIP-1a	PTH	RANK	TGF-b1	IL-17F	Leptin	LRP-6	Lumican	Matrilin-3
TLR3	TLR4	TNFa	TRANCE	TREM-2	MBL	MMP-13A	NOV	sFRP-3	Thrombospondin-5

Data Pre-Processing: Once processed, each array was loaded into Genepix 4000B Microarray scanner (Molecular Devices) and imaged using the "GenePix® Pro 6 Microarray Acquisition & Analysis Software". Proteins for each array type (Table 7.2) were analyzed in quadruplicate. The florescent signal intensity values from 'F532 Mean – B532' for a given protein were included if the value of the spot did not exceed 1.2x the median value of the collective spots for a protein.[107] The average number of replicates included for the Inflammation, immune response, cartilage, and osteoclast arrays across the entire dataset were 3.5, 3.4, 3.6, 3.5 respectively (max is n=4). All spots which met the inclusion criteria were averaged and a log2 transform was applied. Final values were normalized (equation 1) to the average of two control samples (Human Serum Cellect™,

Fisher BioReagents, BP2657-100) run on opposite ends of each array. All data processing took place in Spyder IDE using Python 3.12 and Microsoft Excel.

Equation 1: $$Normalized\ Samples = \frac{Averaged\ Raw\ Sample}{Averaged\ Raw\ Control}$$

Statistical Analysis: Samples were split into cohorts based on the origin of the BMAC samples (iliac crest or humeral head), patient sex (male or female), and patient smoking status (current, former, or never-smoker). To assess normality the Shapiro-Wilk test was used. Two-group sample comparisons were performed using the non-parametric Mann Whitney U test and three-group sample comparisons used a Kruskal-Wallis test with Dunn's post hoc analysis. All samples were subject to false discovery corrections using the Benjamini-Hochberg method to control multiple statistical testing. For all tests a significance threshold of *p* < 0.05 was used.

7.4 Results: A statistical analysis was conducted on the BMAC dataset comprised of 120 patient observations screening 109 proteins per patient. The BMAC dataset was initially stratified by male (n = 55) and female (n = 65) patients. The normality of each protein in the dataset (n = 109) was assessed using a Shapiro-Wilk test where a p-value < 0.05 would suggest a given protein does not have a normal distribution. Given that more than half of the proteins in the dataset test as having a non-normal distribution a Mann-Whitney U (MWU) test was used as a non-parametric two-group comparison test. Benjamini-Hochberg false discovery rate (FDR) corrections were applied to the raw p-values obtained from the MWU testing to control falsely identified significant

proteins (Type 1 error). Of the 109 proteins examined, only three proteins being BMP-5, BMPR-IA, and Matrilin-3 had a raw p-value less than 0.05, however when FDR corrections were applied those proteins were no longer statistically. This process was repeated with the male-female groups, except, each group was further stratified by their BMAC origin site (humeral head or iliac crest). The results can be seen in Table 7.3.

Table 7.3 Statistical analysis for BMAC data set patients stratified by Male (M) and Female (F) with respect to sample extraction site (Humeral Head, Iliac Crest, Both): *Shapiro-Wilk test, Mann-Whitney U (MWU) test, and Benjamini-Hochberg false discovery rate corrections for MWU results as well as significant proteins (raw P < 0.05) are listed in columns 3,4,5, and 6 respectively.*

Dataset	Group Size (M / F)	Shapiro Non-Normal (M / F)	MWU Raw / FDR Sig Proteins (p < 0.05)	Significant Proteins (Raw p < 0.05)
Iliac Crest (77)	31 M / 46 F	56 / 76 → 49 overlap	11/0	BMP-5, BMPR-IA, FGF-4, FGF-6, Matrilin-3, Thrombospondin-5, bIG-H3, CRP, CD109, Ephrin-A4, IL-20
Humeral Head (43)	24 M / 19 F	51 / 63 → 33 overlap	0/0	—
Both (120)	55 M / 65 F	77 / 89 → 70 overlap	3/0	BMP-5, BMPR-IA, Matrilin-3

The BMAC dataset was stratified by patient smoking status being, smoker (n = 24), former smoker (n = 31), or never smoker (n = 65). Again, the Shapiro-

Wilk test was used to assess the normality of each group ($p<0.05$); for 109 total proteins, the smoker group had 51, former smokers had 65, and never smokers had 85 proteins with a non-normal distribution. The Kruskal-Wallis (KW) test was used as a non-parametric multi-group ($2<n$) comparison test to identify any significant differences between the smoker, former smoker, and the never smoker groups. The KW test identified BMPR-II, G-CSF, IL-12p70, IL-17, and TRANCE as statistically significantly different within the test groups. A Dunn's post-hoc test was applied as a pairwise comparison to identify the location of the significant proteins identified by the KW test. False discovery rate (FDR) correction using the Benjamini–Hochberg method was applied to both the Kruskal-Wallis p-values (across all proteins) and the Dunn post hoc p-values (within each protein's pairwise comparisons) to control for multiple hypothesis testing. This procedure was performed twice more on the dataset, however, the BMAC extraction site (humeral head or iliac crest) was taken into account. The results are summarized in Table 7.4.

Table 7.4. Statistical analysis for BMAC data set patients stratified by Smoking status with respect to sample extraction site (Humeral Head, Iliac Crest, Both) *Summarized results for all Kruskal-Wallis and Dunn's post hoc statistical analysis. All proteins reported had a 'Raw P-value' of p < 0.05. A false discovery rate (FDR) correction using the Benjamini-Hochberg method; All proteins whose 'FDR P-value' was p < 0.05, were statistically significant. The following three columns: FS vs N', FS vs S', and 'N vs S' were the result of Dunn's post hoc test, values listed in the columns are p-values where p <0.05 was statistically significant.*
Note that FS = Former Smokers, S = Smokers, N = Never smokers.

Site	Protein	Kruskal H-statistic	Raw P-value	FDR P-value	FS vs N (p)	FS vs S (p)	N vs S (p)
Both	BMPR-II	6.3119	0.0426	0.2512	**0.039**	0.291	0.291
Both	G-CSF	7.3207	0.0257	0.2512	0.21	0.272	**0.028**
Both	IL-12p70	6.4556	0.0396	0.2512	0.089	0.689	0.088
Both	IL-17	11.7246	0.0028	0.1526	**0.004**	0.416	0.064
Both	TRANCE	12.9215	0.0016	0.1526	0.098	0.098	**0.001**
HH	Aggrecan	6.6135	0.0366	0.7495	0.072	0.781	0.072
HH	TGF-b1	6.2916	0.043	0.7495	0.958	0.051	0.051
HH	TLR3	6.0156	0.0494	0.7495	0.577	0.052	0.052
HH	TREM-2	8.1685	0.0168	0.7495	0.233	**0.014**	0.051
HH	VCAM-1	6.3238	0.0423	0.7495	0.192	**0.036**	0.147
IC	BLC	9.4233	0.009	**0.0381**	0.289	0.095	**0.007**
IC	BMP-9	11.1845	0.0037	**0.0192**	**0.016**	0.656	**0.015**
IC	BMPR-II	11.1974	0.0037	**0.0192**	**0.019**	0.598	**0.013**
IC	Eotaxin	12.902	0.0016	**0.0138**	0.078	0.133	**0.002**
IC	Ephrin-A4	13.7183	0.001	**0.0121**	0.064	0.128	**0.001**
IC	E-Selectin	7.5057	0.0235	0.0745	0.171	0.301	**0.028**
IC	Fas	6.8304	0.0329	0.0944	0.337	**0.031**	0.061
IC	FGF-9	9.3909	0.0091	**0.0381**	0.082	0.33	**0.013**
IC	FSH	11.746	0.0028	**0.017**	**0.033**	0.363	**0.006**
IC	G-CSF	10.583	0.005	**0.0237**	0.123	0.123	**0.004**
IC	GM-CSF	14.227	0.0008	**0.0121**	**0.036**	0.183	**0.001**
IC	ICAM-1	7.4576	0.024	0.0745	0.761	0.051	**0.022**
IC	IFNg	19.9175	0	**0**	**0.002**	0.388	**0.0003**
IC	IL-10	6.6769	0.0355	0.0992	0.203	0.331	**0.043**
IC	IL-11	15.285	0.0005	**0.0109**	**0.019**	0.234	**0.001**
IC	IL-12p70	14.0289	0.0009	**0.0121**	**0.023**	0.272	**0.002**
IC	IL-13	6.0628	0.0482	0.1251	0.268	0.268	**0.046**
IC	IL-15	12.6934	0.0018	**0.0138**	**0.039**	0.255	**0.003**
IC	IL-16	7.4124	0.0246	0.0745	0.087	0.531	**0.046**
IC	IL-17	14.7277	0.0006	**0.0109**	**0.007**	0.527	**0.003**
IC	IL-18	10.593	0.005	**0.0237**	0.948	**0.008**	**0.006**
IC	IL-1a	12.3737	0.0021	**0.0141**	**0.046**	0.239	**0.003**
IC	IL-1ra	8.7979	0.0123	**0.0462**	0.269	0.127	**0.009**
IC	IL-2	12.8002	0.0017	**0.0138**	**0.027**	0.321	**0.003**
IC	IL-2 Ra	6.8912	0.0319	0.094	0.934	**0.045**	**0.034**
IC	IL-23	12.6293	0.0018	**0.0138**	**0.007**	0.752	**0.007**
IC	IL-4	8.9643	0.0113	**0.044**	0.239	0.138	**0.009**
IC	IL-5	15.0417	0.0005	**0.0109**	**0.016**	0.278	**0.001**
IC	IL-7	15.2697	0.0005	**0.0109**	**0.009**	0.4	**0.002**
IC	Matrilin-3	7.6288	0.0221	0.0733	0.218	0.233	**0.023**
IC	MCP-1	8.9847	0.0112	**0.044**	0.06	0.461	**0.021**
IC	MCSF	12.2665	0.0022	**0.0141**	**0.009**	0.873	**0.009**
IC	MIP-1a	8.4892	0.0143	0.052	0.196	0.196	**0.012**
IC	RANK	8.0848	0.0176	0.0619	**0.036**	0.815	**0.036**
IC	sFRP-3	6.3755	0.0413	0.1098	0.077	0.838	0.112
IC	TGF-b1	12.9576	0.0015	**0.0138**	**0.031**	0.285	**0.003**
IC	Thrombospondin-5	6.5415	0.038	0.1036	0.477	0.138	**0.032**
IC	TIMP-2	14.7166	0.0006	**0.0109**	**0.008**	0.461	**0.002**
IC	TNFa	12.5424	0.0019	**0.0138**	**0.03**	0.316	**0.004**
IC	TNFb	11.5443	0.0031	**0.0178**	0.059	0.244	**0.004**
IC	TRANCE	9.9288	0.007	**0.0318**	0.144	0.175	**0.007**

Following the statistical analysis proteins which had a raw P-value of $p < 0.05$ (listed in Table 7.3) were plotted by their normalized RFU's with respect to either age, BMI, or sample density as seen in the following figures 7.1 through 7.4. Note that only proteins whose expression levels were significantly different between 'smokers and non-smokers' AND 'never-smokers and former smokers' after FDR corrections, were included in these figures.

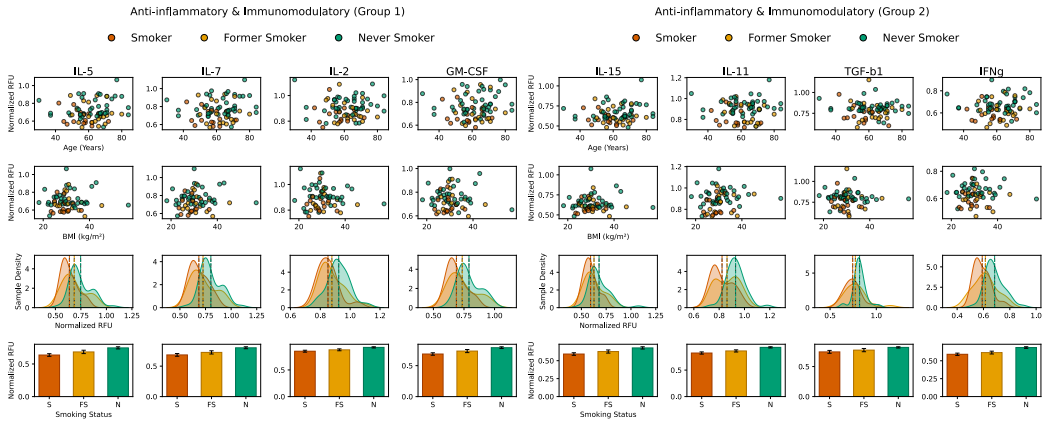

Figure 7.1 Anti-inflammatory and Immunomodulatory proteins: Visualization of the relative normalized fluorescent units (RFU) for each protein: *For each protein column (i.e., protein), the subsequent rows depict*
(a) Scatter plot of normalized RFU vs. patient age,
(b) Scatter plot of normalized RFU vs. patient BMI,
(c) Kernel density plot showing distribution of RFU across smoking cohorts, with dashed vertical lines indicating group means,
(d) Bar plot of average RFU per smoking status with standard error bars.
Proteins are grouped arbitrarily into sets (e.g., Group 1, Group 2) based on panel layout for figure clarity; these groupings do not represent biological or statistical cohorts.

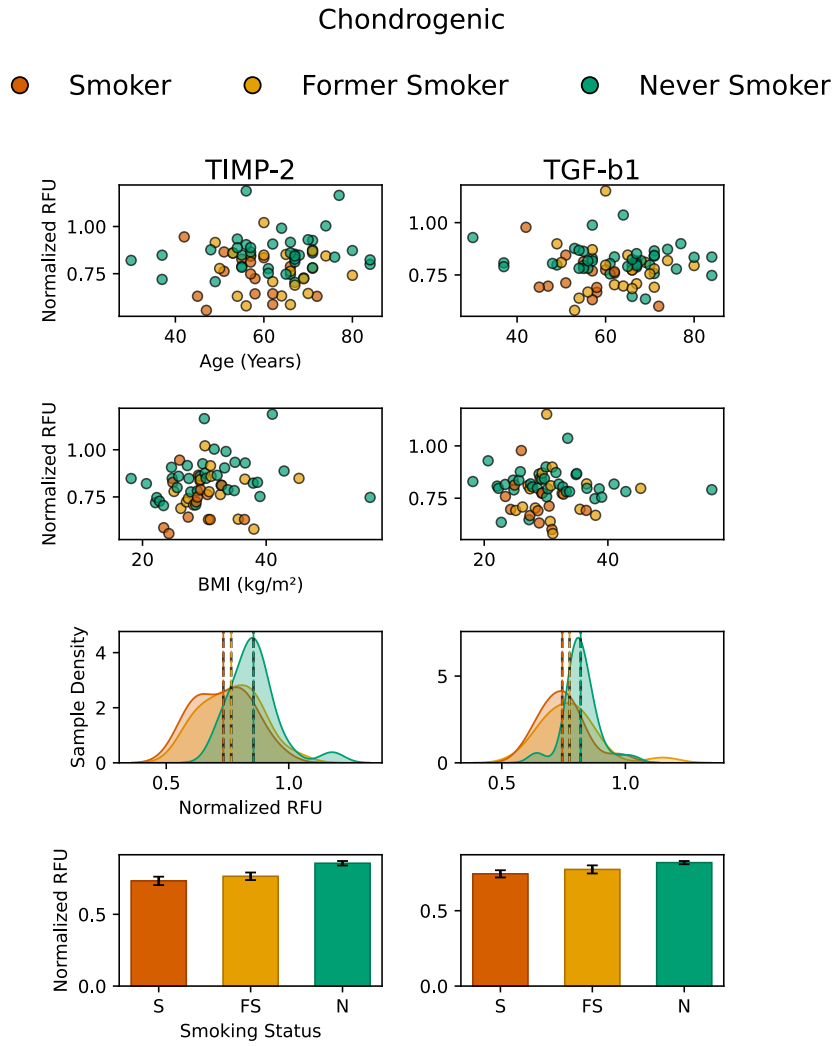

Figure 7.2 Chondrogenic proteins: Visualization of the relative normalized fluorescent units (RFU) for each protein: *For each protein column (i.e., protein), the subsequent rows depict*
(a) Scatter plot of normalized RFU vs. patient age,
(b) Scatter plot of normalized RFU vs. patient BMI,
(c) Kernel density plot showing distribution of RFU across smoking cohorts, with dashed vertical lines indicating group means,
(d) Bar plot of average RFU per smoking status with standard error bars.
Proteins are grouped arbitrarily into sets (e.g., Group 1, Group 2) based on panel layout for figure clarity; these groupings do not represent biological or statistical cohorts.

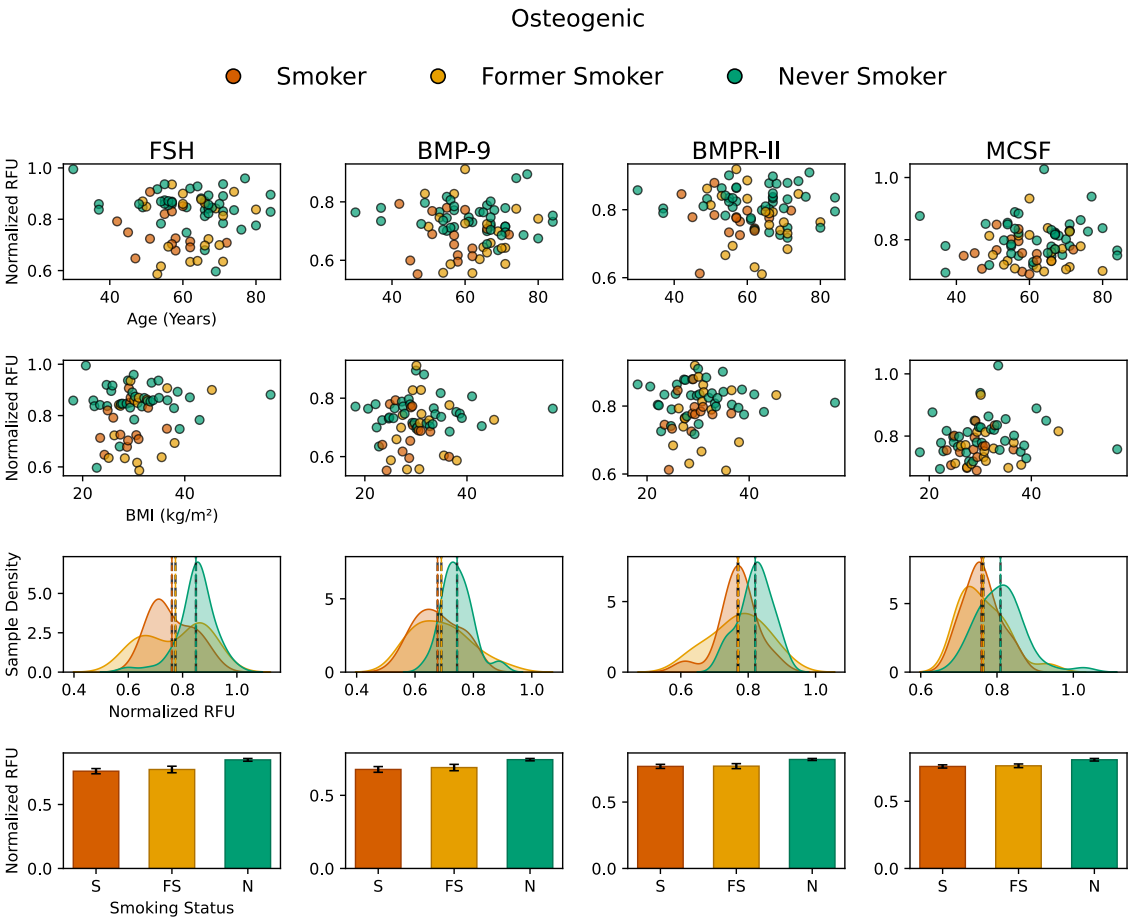

Figure 7.3 Osteogenic proteins: Visualization of the relative normalized fluorescent units (RFU) for each protein: *For each protein column (i.e., protein), the subsequent rows depict*
(a) Scatter plot of normalized RFU vs. patient age,
(b) Scatter plot of normalized RFU vs. patient BMI,
(c) Kernel density plot showing distribution of RFU across smoking cohorts, with dashed vertical lines indicating group means,
(d) Bar plot of average RFU per smoking status with standard error bars.
Proteins are grouped arbitrarily into sets (e.g., Group 1, Group 2) based on panel layout for figure clarity; these groupings do not represent biological or statistical cohorts.

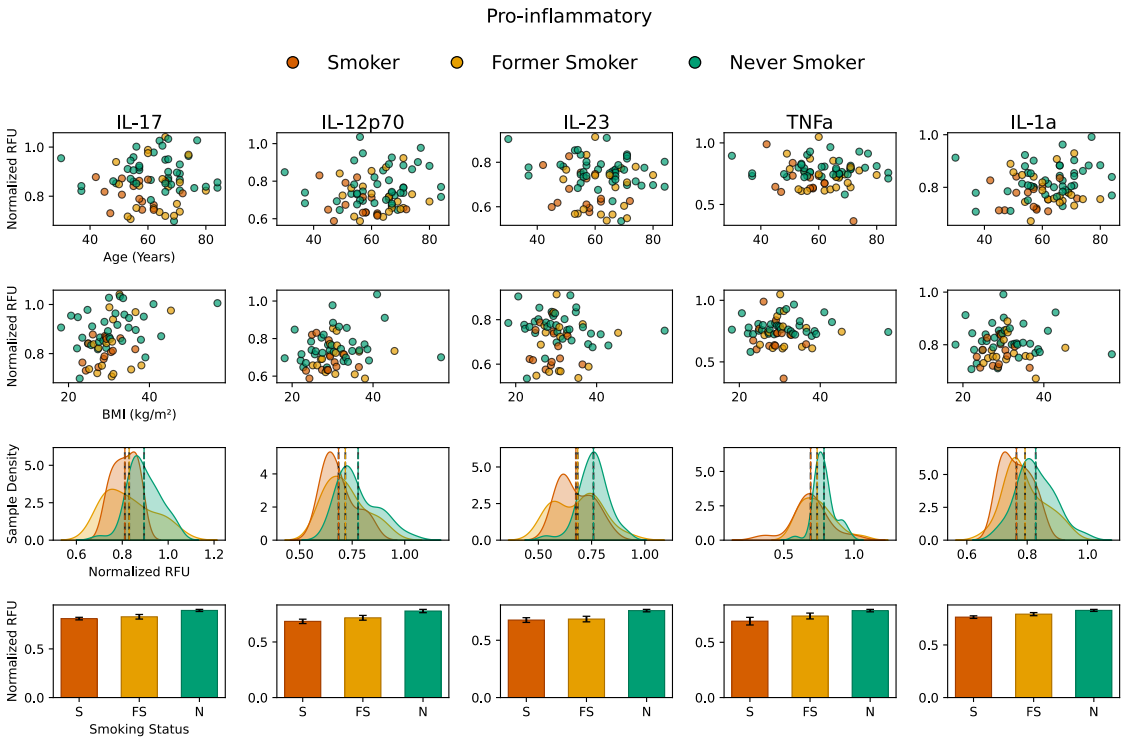

Figure 7.4 Pro-Inflammatory proteins: Visualization of the relative normalized fluorescent units (RFU) for each protein: *For each protein column (i.e., protein), the subsequent rows depict*
(a) Scatter plot of normalized RFU vs. patient age,
(b) Scatter plot of normalized RFU vs. patient BMI,
(c) Kernel density plot showing distribution of RFU across smoking cohorts, with dashed vertical lines indicating group means,
(d) Bar plot of average RFU per smoking status with standard error bars.
Proteins are grouped arbitrarily into sets (e.g., Group 1, Group 2) based on panel layout for figure clarity; these groupings do not represent biological or statistical cohorts.

7.5 Discussion:

The BMAC dataset is currently composed of 120 patient observations, each described by 109 unique proteins; these proteins were analyzed using one of four antibody microarrays (Table 7.2). The choice for using the inflammation, immune response, osteoclast, and chondrogenic arrays largely comes from their analyte diversity while also encompassing many proteins that other works may describe as integral to the nature of BMAC's regenerative properties. Notable examples include, BMP-2, BMP-7, IL-1ra, TGF-B1, VEGF, and IL-8.[149,177]

Although other arrays could have been informative in isolation, our selection captures many of the canonical pathways associated with BMAC. Notably, the design of this study allows for future horizontal expansion (additional proteins) and vertical expansion (additional patient observations).

Stated previously, the role of patient demographic factors is largely unstudied in the context of BMAC. Our goal was to identify if and how a given patients age, sex, BMI, or smoking status may play a role in the proteomic composition of BMAC.

Our analyses looked at the BMAC samples in one of two ways, stratified by bone marrow source or grouped as a whole. The rationale for bone marrow source stratification predominately comes from our previous work which identified several key differences between the BMAC proteomes depending on the location of the bone marrows extraction.[54] Further stratifications were used to analyze patients either by biological sex or smoking status.

A Shapiro-Wilk's test was used to assess the normality of each protein being analyzed with respect to each division of groups. As expected, many of the proteins failed the Shapiro-Wilk test, obtaining a p-value of $p < 0.05$, meaning the protein in question lacked a normal distribution. This information is important in the way of determining what types of statistical tests are appropriate for the dataset. Given the large mixture of normally and non-normally distributed proteins, we opted for a non-parametric series of statistical tests (Table 7.3 and 7.4).[115,116]

We first examined patients by their biological sex. The normality of each group was mixed and therefore a Mann-Whitney U (MWU) test was used to perform two-group comparison. The unadjusted or raw MWU results showed that 11 proteins were identified as statistically significantly different between males and females whose BMAC sample originated from the iliac crest, 0 proteins for the humeral head samples, and 3 proteins with all samples included (humeral head + iliac crest). In total 11 unique proteins were characterized as statistically significant these proteins being: BMP-5, BMPR-IA, FGF-4, FGF-6, Matrilin-3, Thrombospondin-5, bIG-H3, CRP, CD109, Ephrin-A4, and IL-20.
Notably, none of these proteins survived the Benjamini-Hochberg FDR corrections, all proteins having an adjusted p-value of $p > 0.05$. Contextually, FDR corrections are often used in -omics as a means of controlling falsely significant features (proteins, genes, biomarkers, etc). However, given the smaller sample size of our dataset it is important to take protein functions into account when determining the magnitude of their effect on a system.

Our results suggest that biological sex differences do manifest in the BMAC proteome albeit in a subset of proteins whose functions directly relate to bone and cartilage development in the way of BMPs, FGFs, and extracellular matrix proteins like Matrilin-3, Thrombospondin-5, and bIG-H3.[154,178–181]

To analyze the smoking cohorts, we stratified patients by bone marrow source (iliac crest, humeral head, or both) in addition to their smoking statuses (current, former, or never smokers). The normality of each protein (Table 7.2) within the cohorts was assessed using the Shapiro-Wilk test. Following this, a Kruskal-Wallis' (KW) test was performed as the non-parametric equivalent to an ANVOA; this test enabled us to identify that there are differences between $n > 2$ groups of samples and produces an H-statistic which relates to how big those differences are. The KW test identified 51 proteins across 3 testing groups which had a raw p-value of $p < 0.05$. When Benjamini-Hochberg FDR corrections were applied to these results, only 29 of those proteins survived; All proteins which contained a significant FDR adjusted p-value where $p < 0.05$ belonged to the iliac crest smoking cohort (Table 7.4). Within these 29 proteins, we were able to use Dunn's post hoc testing (with FDR corrections) to identify where the differences were between smokers, former smokers, and never smokers.

Smoking status was found to profoundly influence the BMAC proteome, with 29 proteins in the iliac crest cohort showing FDR-corrected significant p-values. Of the 29 proteins, 18 were differentially expressed between smokers and never smokers, with former smokers often exhibiting moderate profiles. Notably smoking patients from the iliac crest cohort demonstrated a pro-

inflammatory skew, having elevated levels of catabolic mediators (e.g., IL-17, TNF-a, IL-1a) and reduced anti-inflammatory proteins (e.g., IL-1ra, IL-4). Osteo- and chondrogenic factors such as BMP-9, FGF-9, TGF-B1 showed lower levels in the smoking patients, suggesting impaired regeneration (Tables 4 and 5).
It is important to note the interplay between augmented levels of proteins in the bone marrow. For example, IL-1ra being down-regulated in smoking patients while IL-1a is simultaneously up-regulated may contribute to the catabolic environment. Should a protein like TRANCE (RANKL) also be upregulated, this could promote greater levels of osteoclast activity, which may mirror smoking-associated bone loss. Likewise, reduced expression of proteins like TGF-B1 and BMP-9 would align with studies which link cigarette smoke exposure (CSE) to impair osteogenesis.[174] CSE has also been shown in mice to reduce chondrocyte viability, proliferation, and matrix formation in both a dose and time dependent manner.[182]

Of particular interest is the site-specific effects which smoking appears to have; Based on our statistical analysis, the iliac crest cohort appears to have a greater degree of variance between patent smoking status'. We found the lack of differentiation between smoking status in the humeral head cohort surprising for a few reasons; Namely, there is link between a patient smoking and rotator cuff injury.[183–186] Through a combination of fatty infiltration[187,188] of the rotator cuff based combined with the plethora of fat-soluble compounds in cigarette smoke[189], we had expected a larger observable difference between patient in the humeral head cohort than actually observed.

We acknowledge that with larger sample sets (especially in the humeral head cohort) the explanatory power of our current results could be greatly improved. Additionally, one of the largest limitations of this work is not having access to more details around the former smokers (in all cohorts), we do not have access to the duration of time patients stopped smoking, nor the volume at which they were smoking; Future studies should seek to stratify their patients using those metrics when possible because there is a very strong possibility that the differential protein expression between former-smokers and the other groups may have been far more pronounced. Additionally, while the modulation of proteins in isolation are apparent with respect to the bone marrow site and smoking status, the interactions between multiple modulated proteins would require *in vitro* validation for a true causal pathway.

Table 7.5. Tabulation of all proteins whose FDR-adjusted p-value (listed in Table 4) was p < 0.05: *All proteins listed contain their name, general function, and classification. Note that all proteins in this table are from the Iliac Crest Kruskal-Wallis testing group. Items preceded by * = Different between Smokers and Never Smokers only, ** = Different between Smokers and Never Smokers as well as Former Smokers and Smokers. Items with no prefix were different between both the Smokers and Never Smokers as well as Former Smokers and Never Smokers.*

Protein	Function	Classification	Reference
*IL-1ra	Antagonist of IL-1, anti-inflammatory	Anti-inflammatory	4,29
*IL-4	Anti-inflammatory promotes Th2 responses	Anti-inflammatory, Immunomodulatory	149
TIMP-2	Inhibits MMPs, regulates extracellular matrix turnover	Chondrogenic	177
*FGF-9	Promotes chondrocyte proliferation	Chondrogenic	177,179
TGF-β1	Anti-inflammatory, promotes chondrogenesis and fibrosis	Chondrogenic, Anti-inflammatory	29,149,164,178
IL-5	Stimulates eosinophils, associated with allergic responses	Immunomodulatory	149
IL-7	Lymphocyte development and survival	Immunomodulatory	149
IL-2	T-cell growth factor, promotes immune activation	Immunomodulatory	149
*G-CSF	Stimulates neutrophil production, immunomodulatory	Immunomodulatory	149
*BLC (CXCL13)	B-cell chemokine, involved in lymphoid organization	Immunomodulatory	149
*Ephrin-A4	Cell signaling, angiogenesis, bone remodeling	Osteogenic	149,162
FSH	Hormone with potential bone resorption effects	Osteogenic	149
BMP-9	Promotes osteoblast differentiation	Osteogenic	178
BMPR-II	Receptor for BMPs, involved in osteogenesis	Osteogenic	178
*TRANCE (RANKL)	Essential for osteoclast differentiation	Osteogenic	149,162
IL-11	Promotes osteoblast differentiation, anti-inflammatory in some contexts	Osteogenic, Immunomodulatory	149
M-CSF	Stimulates osteoclast formation, regulates macrophage function	Osteogenic, Immunomodulatory	149
IL-17	Pro-inflammatory cytokine, promotes bone resorption, Th17 responses	Pro-inflammatory	149
IL-12p70	Promotes Th1 responses, pro-inflammatory	Pro-inflammatory	149
*Eotaxin	Chemokine for eosinophils, involved in allergic inflammation	Pro-inflammatory	149
IL-23	Promotes Th17 responses, pro-inflammatory	Pro-inflammatory	149
TNF-α	Pro-inflammatory, promotes bone resorption	Pro-inflammatory	4,149
IL-1α	Pro-inflammatory, promotes cartilage degradation	Pro-inflammatory	4,149
*TNF-β	Pro-inflammatory, similar to TNF-α	Pro-inflammatory	149
**IL-18	Pro-inflammatory, promotes IFN-γ production	Pro-inflammatory	149
*MCP-1	Recruits monocytes, pro-inflammatory	Pro-inflammatory	149,163
IFN-γ	Pro-inflammatory cytokine, promotes Th1 responses, inhibits osteogenesis	Pro-inflammatory, Immunomodulatory	149,163
GM-CSF	Stimulates myeloid cell proliferation, pro-inflammatory	Pro-inflammatory, Immunomodulatory	149,163
IL-15	Stimulates NK and T cells, pro-inflammatory	Pro-inflammatory, Immunomodulatory	149

7.6 Conclusion:

This study demonstrates that patient demographic factors, namely smoking, significantly altered the proteomic profile of BMAC. We observed the largest differences in the iliac crest cohort of patients, but BMAC samples originating from the iliac crest and humeral head were profiled. In the iliac crest cohort, anabolic growth factors including TGF-B1 and BMP-9 were downregulated and pro-inflammatory cytokines (IL-17, TNF-a) were upregulated; these findings ultimately suggest that smoking undermines the therapeutic properties of BMAC. Former smokers exhibited moderate proteomic profiles that show signs that cessation could restore some of the bone marrow proteome; More information about patient smoking habits and timelines would be required to draw impactful conclusions. Overall, our study's findings point to the need for patient stratification by smoking status in clinical BMAC applications, pre-treatment smoking cessation to optimize patient outcomes, and mechanistic studies to clarify the links between smoking status and BMAC's proteomic changes.

Chapter 8

Future Direction

8.1 Concluding remarks for the future of this work

The findings of this work highlight both the biological complexity and analytical potential inherent in the study of bone marrow aspirate concentrate (BMAC). While this project established a proteomic and computational framework capable of distinguishing site- and demographic-specific protein expression patterns, several key avenues remain open for further exploration. Future efforts should focus on three interconnected domains: data expansion, biological contextualization, and clinical translation.

Data Expansion and Refinement: The most immediate and actionable step is to expand the patient dataset in both size and scope. A larger cohort would increase statistical power, ensure the reproducibility of site- and cohort-specific findings. Additionally, this would enable the use of more sophisticated multivariate models such as ensemble learning or Bayesian hierarchical modeling, to better capture the continuous, non-linear effects of demographic variables on protein expression. Critically, expanding the scope of patient metadata is equally important. For instance, refining the smoking variable to include duration of use and time since cessation would allow for a more precise classification of "former" versus "current" smokers, moving beyond the binary or

limited categorizations that currently constrain analysis. Incorporating additional demographic dimensions, such as race and ethnicity, would further enhance the granularity of analyses like principal component analysis, enabling the identification of more nuanced, subgroup-specific proteomic signatures.

Biological Contextualization: Beyond scaling the dataset, future work must seek to contextualize the identified proteomic signatures within functional biological pathways. Integration of systems biology tools, such as network enrichment analysis, pathway mapping, or multi-omics fusion with transcriptomic or metabolomic data—would allow for a more mechanistic understanding of how demographic and anatomical factors influence the therapeutic potential of BMAC. Parallel validation using targeted immunoassays or targeted mass spectrometry would provide an essential layer of biochemical verification for the biomarkers and trends revealed through machine learning.

Clinical Translation: Finally, translating these computational insights into actionable clinical workflows remains a paramount objective. This entails the development of standardized data acquisition pipelines, robust normalization procedures, and accessible analytical tools. Such resources would empower clinicians and researchers to assess BMAC composition and predict therapeutic potential. By transforming BMAC from a descriptive biologic into a measurable, optimized therapeutic platform, future studies can directly advance the broader goal of personalized regenerative medicine through data-informed design.

Bibliography:

(1) Long, H.; Liu, Q.; Yin, H.; Wang, K.; Diao, N.; Zhang, Y.; Lin, J.; Guo, A. Prevalence Trends of Site-Specific Osteoarthritis From 1990 to 2019: Findings From the Global Burden of Disease Study 2019. *Arthritis Rheumatol.* **2022**, *74* (7), 1172–1183. https://doi.org/10.1002/art.42089.

(2) *GBD Results*. Institute for Health Metrics and Evaluation. https://vizhub.healthdata.org/gbd-results (accessed 2025-10-01).

(3) Fortier, L.; Strauss, E.; Shepard, D.; Becktell, L.; Kennedy, J. Biological Effects of Bone Marrow Concentrate in Knee Pathologies. *J. Knee Surg.* **2019**, *32* (01), 002–008. https://doi.org/10.1055/s-0038-1676069.

(4) Cassano, J. M.; Kennedy, J. G.; Ross, K. A.; Fraser, E. J.; Goodale, M. B.; Fortier, L. A. Bone Marrow Concentrate and Platelet-Rich Plasma Differ in Cell Distribution and Interleukin 1 Receptor Antagonist Protein Concentration. *Knee Surg. Sports Traumatol. Arthrosc.* **2018**, *26* (1), 333–342. https://doi.org/10.1007/s00167-016-3981-9.

(5) Dinarello, C. Biologic Basis for Interleukin-1 in Disease. *Blood* **1996**, *87* (6), 2095–2147. https://doi.org/10.1182/blood.V87.6.2095.bloodjournal8762095.

(6) Arend, W. P.; Malyak, M.; Guthridge, C. J.; Gabay, C. INTERLEUKIN-1 RECEPTOR ANTAGONIST: Role in Biology. *Annu. Rev. Immunol.* **1998**, *16* (1), 27–55. https://doi.org/10.1146/annurev.immunol.16.1.27.

(7) Kaneko, N.; Kurata, M.; Yamamoto, T.; Morikawa, S.; Masumoto, J. The Role of Interleukin-1 in General Pathology. *Inflamm. Regen.* **2019**, *39* (1), 12. https://doi.org/10.1186/s41232-019-0101-5.

(8) Gowd, A. K.; Cabarcas, B. C.; Frank, R. M.; Cole, B. J. Biologic Augmentation of Rotator Cuff Repair: The Role of Platelet-Rich Plasma and Bone Marrow Aspirate Concentrate. *Oper. Tech. Sports Med.* **2018**, *26* (1), 48–57. https://doi.org/10.1053/j.otsm.2017.10.008.

(9) Rumbaut, R. E.; Thiagarajan, P. General Characteristics of Platelets. In *Platelet-Vessel Wall Interactions in Hemostasis and Thrombosis*; Morgan & Claypool Life Sciences, 2010.

(10) Chen, P.-H.; Chen, X.; He, X. Platelet-Derived Growth Factors and Their Receptors: Structural and Functional Perspectives. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **2013**, *1834* (10), 2176–2186. https://doi.org/10.1016/j.bbapap.2012.10.015.

(11) Heldin, C.-H.; Westermark, B. Mechanism of Action and In Vivo Role of Platelet-Derived Growth Factor. *Physiol. Rev.* **1999**, *79* (4), 1283–1316. https://doi.org/10.1152/physrev.1999.79.4.1283.

(12) Ferrara, N. Vascular Endothelial Growth Factor: Basic Science and Clinical Progress. *Endocr. Rev.* **2004**, *25* (4), 581–611. https://doi.org/10.1210/er.2003-0027.

(13) Kubiczkova, L.; Sedlarikova, L.; Hajek, R.; Sevcikova, S. TGF-β – an Excellent Servant but a Bad Master. *J. Transl. Med.* **2012**, *10* (1), 183. https://doi.org/10.1186/1479-5876-10-183.
(14) Bikfalvi, A.; Klein, S.; Pintucci, G.; Rifkin, D. B. Biological Roles of Fibroblast Growth Factor-2.
(15) Benington, L.; Rajan, G.; Locher, C.; Lim, L. Y. Fibroblast Growth Factor 2—A Review of Stabilisation Approaches for Clinical Applications. *Pharmaceutics* **2020**, *12* (6), 508. https://doi.org/10.3390/pharmaceutics12060508.
(16) Laron, Z. Insulin-like Growth Factor 1 (IGF-1): A Growth Hormone. *Mol. Pathol.* **2001**, *54* (5), 311–316. https://doi.org/10.1136/mp.54.5.311.
(17) Wang, Y.; Menendez, A.; Fong, C.; ElAlieh, H. Z.; Kubota, T.; Long, R.; Bikle, D. D. IGF-I Signaling in Osterix-Expressing Cells Regulates Secondary Ossification Center Formation, Growth Plate Maturation, and Metaphyseal Formation During Postnatal Bone Development. *J. Bone Miner. Res.* **2015**, *30* (12), 2239–2248. https://doi.org/10.1002/jbmr.2563.
(18) Berlanga-Acosta, J.; Gavilondo-Cowley, J.; López-Saura, P.; González-López, T.; Castro-Santana, M. D.; López-Mola, E.; Guillén-Nieto, G.; Herrera-Martinez, L. Epidermal Growth Factor in Clinical Practice – a Review of Its Biological Actions, Clinical Indications and Safety Implications. *Int. Wound J.* **2009**, *6* (5), 331–346. https://doi.org/10.1111/j.1742-481X.2009.00622.x.
(19) Saeedi, P.; Halabian, R.; Fooladi, A. A. I. A Revealing Review of Mesenchymal Stem Cells Therapy, Clinical Perspectives and Modification Strategies. *Stem Cell Investig.* **2019**, *6*. https://doi.org/10.21037/sci.2019.08.11.
(20) Hao, J.; Li, S.; Shi, X.; Qian, Z.; Sun, Y.; Wang, D.; Zhou, X.; Qu, H.; Hu, S.; Zuo, E.; Zhang, C.; Hou, L.; Wang, Q.; Piao, F. Bone Marrow Mesenchymal Stem Cells Protect against N-Hexane-Induced Neuropathy through Beclin 1-Independent Inhibition of Autophagy. *Sci. Rep.* **2018**, *8*, 4516. https://doi.org/10.1038/s41598-018-22857-x.
(21) Kusadasi, N.; Groeneveld, A. B. J. A Perspective on Mesenchymal Stromal Cell Transplantation in the Treatment of Sepsis. *Shock* **2013**, *40* (5), 352–357. https://doi.org/10.1097/SHK.0000000000000039.
(22) Brozovich, A.; Sinicrope, B. J.; Bauza, G.; Niclot, F. B.; Lintner, D.; Taraballi, F.; McCulloch, P. C. High Variability of Mesenchymal Stem Cells Obtained via Bone Marrow Aspirate Concentrate Compared With Traditional Bone Marrow Aspiration Technique. *Orthop. J. Sports Med.* **2021**, *9* (12), 23259671211058459. https://doi.org/10.1177/23259671211058459.
(23) Caplan, A. I. Mesenchymal Stem Cells. *J. Orthop. Res.* **1991**, *9* (5), 641–650. https://doi.org/10.1002/jor.1100090504.
(24) Friedenstein, A. J.; Deriglasova, U. F.; Kulagina, N. N.; Panasuk, A. F.; Rudakowa, S. F.; Luriá, E. A.; Ruadkow, I. A. Precursors for Fibroblasts in Different Populations of Hematopoietic Cells as Detected by the in Vitro Colony Assay Method. *Exp. Hematol.* **1974**, *2* (2), 83–92.

(25) Mutlu, L.; Hufnagel, D.; Taylor, H. S. The Endometrium as a Source of Mesenchymal Stem Cells for Regenerative Medicine. *Biol. Reprod.* **2015**, *92* (6), 138. https://doi.org/10.1095/biolreprod.114.126771.
(26) Ma, Z.-J.; Yang, J.-J.; Lu, Y.-B.; Liu, Z.-Y.; Wang, X.-X. Mesenchymal Stem Cell-Derived Exosomes: Toward Cell-Free Therapeutic Strategies in Regenerative Medicine. *World J. Stem Cells* **2020**, *12* (8), 814–840. https://doi.org/10.4252/wjsc.v12.i8.814.
(27) Weiss, A. R. R.; Dahlke, M. H. Immunomodulation by Mesenchymal Stem Cells (MSCs): Mechanisms of Action of Living, Apoptotic, and Dead MSCs. *Front. Immunol.* **2019**, *10*.
(28) Dominici, M.; Le Blanc, K.; Mueller, I.; Slaper-Cortenbach, I.; Marini, F. C.; Krause, D. S.; Deans, R. J.; Keating, A.; Prockop, D. J.; Horwitz, E. M. Minimal Criteria for Defining Multipotent Mesenchymal Stromal Cells. The International Society for Cellular Therapy Position Statement. *Cytotherapy* **2006**, *8* (4), 315–317. https://doi.org/10.1080/14653240600855905.
(29) Abreu, S. C.; Rolandsson Enes, S.; Dearborn, J.; Goodwin, M.; Coffey, A.; Borg, Z. D.; dos Santos, C. C.; Wargo, M. J.; Cruz, F. F.; Loi, R.; DeSarno, M.; Ashikaga, T.; Antunes, M. A.; Rocco, P. R. M.; Liu, K. D.; Lee, J.-W.; Matthay, M. A.; McKenna, D. H.; Weiss, D. J. Lung Inflammatory Environments Differentially Alter Mesenchymal Stromal Cell Behavior. *Am. J. Physiol.-Lung Cell. Mol. Physiol.* **2019**, *317* (6), L823–L831. https://doi.org/10.1152/ajplung.00263.2019.
(30) Abu-El-Rub, E.; Khasawneh, R.; Almahasneh, F.; Zegallai, H. Mesenchymal Stem Cells Derived Secretome as an Innovative Cell-Free Therapeutic Approach. *BIOCELL* **2021**, *46* (4), 907–911. https://doi.org/10.32604/biocell.2022.018306.
(31) Halabian, R.; Tehrani, H. A.; Jahanian-Najafabadi, A.; Habibi Roudkenar, M. Lipocalin-2-Mediated Upregulation of Various Antioxidants and Growth Factors Protects Bone Marrow-Derived Mesenchymal Stem Cells against Unfavorable Microenvironments. *Cell Stress Chaperones* **2013**, *18* (6), 785–800. https://doi.org/10.1007/s12192-013-0430-2.
(32) Halabian, R.; Roudkenar, M. H.; Jahanian-Najafabadi, A.; Hosseini, K. M.; Tehrani, H. A. Co-Culture of Bone Marrow-Derived Mesenchymal Stem Cells Overexpressing Lipocalin 2 with HK-2 and HEK293 Cells Protects the Kidney Cells against Cisplatin-Induced Injury. *Cell Biol. Int.* **2015**, *39* (2), 152–163. https://doi.org/10.1002/cbin.10344.
(33) Wechsler, M. E.; Rao, V. V.; Borelli, A. N.; Anseth, K. S. Engineering the MSC Secretome: A Hydrogel Focused Approach. *Adv. Healthc. Mater.* **2021**, *10* (7), 2001948. https://doi.org/10.1002/adhm.202001948.
(34) Gao, Y.; Yuan, Z.; Yuan, X.; Wan, Z.; Yu, Y.; Zhan, Q.; Zhao, Y.; Han, J.; Huang, J.; Xiong, C.; Cai, Q. Bioinspired Porous Microspheres for Sustained Hypoxic Exosomes Release and Vascularized Bone Regeneration. *Bioact. Mater.* **2022**, *14*, 377–388. https://doi.org/10.1016/j.bioactmat.2022.01.041.

(35) Ferreira, J. R.; Teixeira, G. Q.; Santos, S. G.; Barbosa, M. A.; Almeida-Porada, G.; Gonçalves, R. M. Mesenchymal Stromal Cell Secretome: Influencing Therapeutic Potential by Cellular Pre-Conditioning. *Front. Immunol.* **2018**, *9*.
(36) Ahangar, P.; Mills, S. J.; Cowin, A. J. Mesenchymal Stem Cell Secretome as an Emerging Cell-Free Alternative for Improving Wound Repair. *Int. J. Mol. Sci.* **2020**, *21* (19), 7038. https://doi.org/10.3390/ijms21197038.
(37) Asumda, F. Z. Age-Associated Changes in the Ecological Niche: Implications for Mesenchymal Stem Cell Aging. *Stem Cell Res. Ther.* **2013**, *4* (3), 47. https://doi.org/10.1186/scrt197.
(38) Infante, A.; Rodríguez, C. I. Osteogenesis and Aging: Lessons from Mesenchymal Stem Cells. *Stem Cell Res. Ther.* **2018**, *9* (1), 244. https://doi.org/10.1186/s13287-018-0995-x.
(39) Cheng, A.; Choi, D.; Lora, M.; Shum-Tim, D.; Rak, J.; Colmegna, I. Human Multipotent Mesenchymal Stromal Cells Cytokine Priming Promotes RAB27B-Regulated Secretion of Small Extracellular Vesicles with Immunomodulatory Cargo. *Stem Cell Res. Ther.* **2020**, *11* (1), 539. https://doi.org/10.1186/s13287-020-02050-6.
(40) Cavallo, C.; Boffa, A.; De Girolamo, L.; Merli, G.; Kon, E.; Cattini, L.; Santo, E.; Grigolo, B.; Filardo, G. Bone Marrow Aspirate Concentrate Quality Is Affected by Age and Harvest Site. *Knee Surg. Sports Traumatol. Arthrosc.* **2023**, *31* (6), 2140–2151. https://doi.org/10.1007/s00167-022-07153-6.
(41) Hyer, C. F.; Berlet, G. C.; Bussewitz, B. W.; Hankins, T.; Ziegler, H. L.; Philbin, T. M. Quantitative Assessment of the Yield of Osteoblastic Connective Tissue Progenitors in Bone Marrow Aspirate from the Iliac Crest, Tibia, and Calcaneus: *J. Bone Jt. Surg.-Am. Vol.* **2013**, *95* (14), 1312–1316. https://doi.org/10.2106/JBJS.L.01529.
(42) Huddleston, H. P.; Tauro, T.; Credille, K.; Dandu, N.; Hevesi, M.; Chahla, J.; Forsythe, B.; Verma, N.; Yanke, A. B.; Cole, B. J. Patient Demographic Factors Are Not Associated With Mesenchymal Stromal Cell Concentration in Bone Marrow Aspirate Concentrate. *Arthrosc. Sports Med. Rehabil.* **2023**, *5* (3), e559–e567. https://doi.org/10.1016/j.asmr.2023.02.008.
(43) Chahla, J.; Mannava, S.; Cinque, M. E.; Geeslin, A. G.; Codina, D.; LaPrade, R. F. Bone Marrow Aspirate Concentrate Harvesting and Processing Technique. *Arthrosc. Tech.* **2017**, *6* (2), e441–e445. https://doi.org/10.1016/j.eats.2016.10.024.
(44) Mavrogenis, A. F.; Karampikas, V.; Zikopoulos, A.; Sioutis, S.; Mastrokalos, D.; Koulalis, D.; Scarlat, M. M.; Hernigou, P. Orthobiologics: A Review. *Int. Orthop.* **2023**, *47* (7), 1645–1662. https://doi.org/10.1007/s00264-023-05803-z.
(45) Ruoss, S.; Walker, J. T.; Nasamran, C. A.; Fisch, K. M.; Paez, C. J.; Parekh, J. N.; Ball, S. T.; Chen, J. L.; Ahmed, S. S.; Ward, S. R. Strategies to Identify Mesenchymal Stromal Cells in Minimally Manipulated Human Bone Marrow Aspirate Concentrate Lack Consensus. *Am. J. Sports Med.* **2021**, *49* (5), 1313–1322. https://doi.org/10.1177/0363546521993788.

(46) Mazzocca, A. D.; McCarthy, M. B. R.; Chowaniec, D. M.; Cote, M. P.; Arciero, R. A.; Drissi, H. Rapid Isolation of Human Stem Cells (Connective Tissue Progenitor Cells) From the Proximal Humerus During Arthroscopic Rotator Cuff Surgery. *Am. J. Sports Med.* **2010**, *38* (7), 1438–1447. https://doi.org/10.1177/0363546509360924.
(47) Baek, J.-H.; Lee, S. C.; Lee, D. N.; Ahn, H. S.; Nam, C. H. Effectiveness and Complications of Bone Marrow Aspirate Concentrate in Patients with Knee Osteoarthritis of Kellgren–Lawrence Grades II–III. *Medicina (Mex.)* **2024**, *60* (6), 977. https://doi.org/10.3390/medicina60060977.
(48) Varady, N. H.; Cate, G.; Barghi, A.; Jobe, N.; Yakin, D.; Ylanan, R. C.; Arnold, C. A. Positive Early Clinical Outcomes of Bone Marrow Aspirate Concentrate for Osteoarthritis Using a Novel Fenestrated Trocar. *The Knee* **2020**, *27* (5), 1627–1634. https://doi.org/10.1016/j.knee.2020.08.018.
(49) Hernigou, P.; Merouse, G.; Duffiet, P.; Chevalier, N.; Rouard, H. Reduced Levels of Mesenchymal Stem Cells at the Tendon–Bone Interface Tuberosity in Patients with Symptomatic Rotator Cuff Tear. *Int. Orthop.* **2015**, *39* (6), 1219–1225. https://doi.org/10.1007/s00264-015-2724-8.
(50) Ellera Gomes, J. L.; Da Silva, R. C.; Silla, L. M. R.; Abreu, M. R.; Pellanda, R. Conventional Rotator Cuff Repair Complemented by the Aid of Mononuclear Autologous Stem Cells. *Knee Surg. Sports Traumatol. Arthrosc.* **2012**, *20* (2), 373–377. https://doi.org/10.1007/s00167-011-1607-9.
(51) Martin, S. D.; Eberlin, C. T.; Kucharik, M. P.; Cherian, N. J. Harvest and Application of Bone Marrow Aspirate Concentrate to Address Acetabular Chondral Damage During Hip Arthroscopy. *JBJS Essent. Surg. Tech.* **2023**, *13* (2). https://doi.org/10.2106/JBJS.ST.22.00010.
(52) Encinas, R.; Phillips, T.; Hall, S.; Jackson, J. B.; Gonzalez, T. Outcomes and Complications After Bone Marrow Aspirate Harvest From the Calcaneus. *Foot Ankle Orthop.* **2023**, *8* (3), 24730114231194056. https://doi.org/10.1177/24730114231194056.
(53) Ragni, E.; Taiana, M.; Visconte, C.; Kon, E.; Landoni, S.; Colombo, C.; Mangiavini, L.; Peretti, G.; De Girolamo, L. Side-to-side Characterisation of Cellular Content, Soluble Factors and in Vitro Potential on Chondrocytes for Bone Marrow Aspirate Concentrate and Adipose-derived Stromal Vascular Fraction. *J. Exp. Orthop.* **2025**, *12* (2), e70254. https://doi.org/10.1002/jeo2.70254.
(54) Herna, C.; Cipriani, M.; Zhang, Y.; Sachdev, R.; Jedlicka, S. Proteomic Analysis of BMAC from the Iliac Crest and Humeral Head Using Support Vector Machines. *ACS Omega* **2025**, *10* (31), 34975–34985. https://doi.org/10.1021/acsomega.5c03945.
(55) Hu, C.-H.; Tseng, Y.-W.; Chiou, C.-Y.; Lan, K.-C.; Chou, C.-H.; Tai, C.-S.; Huang, H.-D.; Hu, C.-W.; Liao, K.-H.; Chuang, S.-S.; Yang, J.-Y.; Lee, O. K. Bone Marrow Concentrate-Induced Mesenchymal Stem Cell Conditioned Medium Facilitates Wound Healing and Prevents Hypertrophic Scar Formation in a

Rabbit Ear Model. *Stem Cell Res. Ther.* **2019**, *10* (1), 275. https://doi.org/10.1186/s13287-019-1383-x.

(56) Ruoss, S.; Nasamran, C. A.; Ball, S. T.; Chen, J. L.; Halter, K. N.; Bruno, K. A.; Whisenant, T. C.; Parekh, J. N.; Dorn, S. N.; Esparza, M. C.; Bremner, S. N.; Fisch, K. M.; Engler, A. J.; Ward, S. R. Comparative Single-Cell Transcriptional and Proteomic Atlas of Clinical-Grade Injectable Mesenchymal Source Tissues. *Sci. Adv.* **2024**, *10* (28), eadn2831. https://doi.org/10.1126/sciadv.adn2831.

(57) Carlsson, A.; Wingren, C.; Ingvarsson, J.; Ellmark, P.; Baldertorp, B.; Fernö, M.; Olsson, H.; Borrebaeck, C. A. K. Serum Proteome Profiling of Metastatic Breast Cancer Using Recombinant Antibody Microarrays. *Eur. J. Cancer* **2008**, *44* (3), 472–480. https://doi.org/10.1016/j.ejca.2007.11.025.

(58) Ingvarsson, J.; Wingren, C.; Carlsson, A.; Ellmark, P.; Wahren, B.; Engström, G.; Harmenberg, U.; Krogh, M.; Peterson, C.; Borrebaeck, C. A. K. Detection of Pancreatic Cancer Using Antibody Microarray-Based Serum Protein Profiling. *PROTEOMICS* **2008**, *8* (11), 2211–2219. https://doi.org/10.1002/pmic.200701167.

(59) Carlsson, A.; Wuttge, D. M.; Ingvarsson, J.; Bengtsson, A. A.; Sturfelt, G.; Borrebaeck, C. A. K.; Wingren, C. Serum Protein Profiling of Systemic Lupus Erythematosus and Systemic Sclerosis Using Recombinant Antibody Microarrays. *Mol. Cell. Proteomics MCP* **2011**, *10* (5), M110.005033. https://doi.org/10.1074/mcp.M110.005033.

(60) Carlsson, A.; Persson, O.; Ingvarsson, J.; Widegren, B.; Salford, L.; Borrebaeck, C. A. K.; Wingren, C. Plasma Proteome Profiling Reveals Biomarker Patterns Associated with Prognosis and Therapy Selection in Glioblastoma Multiforme Patients. *PROTEOMICS – Clin. Appl.* **2010**, *4* (6–7), 591–602. https://doi.org/10.1002/prca.200900173.

(61) Haab, B. B.; Dunham, M. J.; Brown, P. O. Protein Microarrays for Highly Parallel Detection and Quantitation of Specific Proteins and Antibodies in Complex Solutions. *Genome Biol.* **2001**, *2* (2), research0004.1. https://doi.org/10.1186/gb-2001-2-2-research0004.

(62) Haab, B. B. Methods and Applications of Antibody Microarrays in Cancer Research. *PROTEOMICS* **2003**, *3* (11), 2116–2122. https://doi.org/10.1002/pmic.200300595.

(63) Quintana, F. J.; Farez, M. F.; Viglietta, V.; Iglesias, A. H.; Merbl, Y.; Izquierdo, G.; Lucas, M.; Basso, A. S.; Khoury, S. J.; Lucchinetti, C. F.; Cohen, I. R.; Weiner, H. L. Antigen Microarrays Identify Unique Serum Autoantibody Signatures in Clinical and Pathologic Subtypes of Multiple Sclerosis. *Proc. Natl. Acad. Sci.* **2008**, *105* (48), 18889–18894. https://doi.org/10.1073/pnas.0806310105.

(64) Robinson, W.; Huber, W.; van Venrooij, W.; Muller, S.; Panayi, G.; Verweij, C.; Smolen, J.; Steinman, L.; Utz, P. Protein Microarray Characterization of Autoantibody Responses in Rheumatoid Arthritis and Systemic Lupus Erythematosus. *Arthritis Res.* **2002**, *4* (Suppl 1), 76. https://doi.org/10.1186/ar521.

(65) Kelleher, J. D.; Tierney, B. *Data Science*; The MIT Press essential knowledge series; The MIT Press: Cambridge, Massachusetts London, England, 2018.
(66) Dhar, V. Data Science and Prediction. *Commun. ACM* **2013**, *56* (12), 64–73. https://doi.org/10.1145/2500499.
(67) Ohlsson, M.; Hellmark, T.; Bengtsson, A. A.; Theander, E.; Turesson, C.; Klint, C.; Wingren, C.; Ekstrand, A. I. Proteomic Data Analysis for Differential Profiling of the Autoimmune Diseases SLE, RA, SS, and ANCA-Associated Vasculitis. *J. Proteome Res.* **2021**, *20* (2), 1252–1260. https://doi.org/10.1021/acs.jproteome.0c00657.
(68) Díez, P.; Dasilva, N.; González-González, M.; Matarraz, S.; Casado-Vela, J.; Orfao, A.; Fuentes, M. Data Analysis Strategies for Protein Microarrays. *Microarrays* **2012**, *1* (2), 64–83. https://doi.org/10.3390/microarrays1020064.
(69) Dasu, T.; Johnson, T. *Exploratory Data Mining and Data Cleaning*; Wiley series in probability and statistics; John Wiley & Son: Hoboken, NJ, 2003. https://doi.org/10.1002/0471448354.
(70) Zhu, H.; Snyder, M. Protein Chip Technology. *Curr. Opin. Chem. Biol.* **2003**, *7* (1), 55–63. https://doi.org/10.1016/S1367-5931(02)00005-4.
(71) Hamelinck, D.; Zhou, H.; Li, L.; Verweij, C.; Dillon, D.; Feng, Z.; Costa, J.; Haab, B. B. Optimized Normalization for Antibody Microarrays and Application to Serum-Protein Profiling*S. *Mol. Cell. Proteomics* **2005**, *4* (6), 773–784. https://doi.org/10.1074/mcp.M400180-MCP200.
(72) Li, S.; Song, G.; Bai, Y.; Song, N.; Zhao, J.; Liu, J.; Hu, C. Applications of Protein Microarrays in Biomarker Discovery for Autoimmune Diseases. *Front. Immunol.* **2021**, *12*.
(73) Provost, F.; Fawcett, T. Data Science and Its Relationship to Big Data and Data-Driven Decision Making. *Big Data* **2013**, *1* (1), 51–59. https://doi.org/10.1089/big.2013.1508.
(74) Gigerenzer, G.; Hoffrage, U. How to Improve Bayesian Reasoning without Instruction: Frequency Formats. *Psychol. Rev.* **1995**, *102* (4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684.
(75) Eisenhauer, J. G. The Monty Hall Matrix. *Teach. Stat.* **2000**, *22* (1), 17–20. https://doi.org/10.1111/1467-9639.00005.
(76) Brown, T. A. The Monty Hall Dilemma Donald Granberg. *Pers. Soc. Psychol. Bull.* **1995**, *21* (7), 711–723. https://doi.org/10.1177/0146167295217006.
(77) Grimmett, G. G.; Stirzaker, D. S. *Probability and Random Processes*, 3rd ed.; Oxford university press: Oxford, 2001.
(78) Ross, S. M. *Introduction to Probability and Statistics for Engineers and Scientists*, Sixth edition.; Academic Press, an imprint of Elsevier: London, UK San Diego, CA Cambridge, MA Oxford, UK, 2021.
(79) Kolmogorov, A. N. *Foundations of the Theory of Probability*; New York: Chelsea Pub. Co., 1950.
(80) *Introduction to Bayesian Statistics*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-72726-2.
(81) Gelman, A. *Bayesian Data Analysis*, 3rd ed.; CRC Press, 2013.

(82) Kruschke, J. K. *Doing Bayesian Data Analysis: A Tutorial Introduction with R, JAGS, and Stan*, Edition 2.; Elsevier, Academic Press: Amsterdam Boston Heidelberg London New York Oxford Paris San Diego San Francisco Singapore Sydney Tokyo, 2015.
(83) Fisher, R. On the Mathematical Foundations of Theoretical Statistics. *Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character* **1922**, *222* (594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009.
(84) Cox, D. R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1958**, *20* (2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.
(85) Fornacon-Wood, I.; Mistry, H.; Johnson-Hart, C.; Faivre-Finn, C.; O'Connor, J. P. B.; Price, G. J. Understanding the Differences Between Bayesian and Frequentist Statistics. *Int. J. Radiat. Oncol.* **2022**, *112* (5), 1076–1082. https://doi.org/10.1016/j.ijrobp.2021.12.011.
(86) Wagenmakers, E.-J. A Practical Solution to the Pervasive Problems Ofp Values. *Psychon. Bull. Rev.* **2007**, *14* (5), 779–804. https://doi.org/10.3758/BF03194105.
(87) Polack, F. P.; Thomas, S. J.; Kitchin, N.; Absalon, J.; Gurtman, A.; Lockhart, S.; Perez, J. L.; Pérez Marc, G.; Moreira, E. D.; Zerbini, C.; Bailey, R.; Swanson, K. A.; Roychoudhury, S.; Koury, K.; Li, P.; Kalina, W. V.; Cooper, D.; Frenck, R. W.; Hammitt, L. L.; Türeci, Ö.; Nell, H.; Schaefer, A.; Ünal, S.; Tresnan, D. B.; Mather, S.; Dormitzer, P. R.; Şahin, U.; Jansen, K. U.; Gruber, W. C. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N. Engl. J. Med.* **2020**, *383* (27), 2603–2615. https://doi.org/10.1056/NEJMoa2034577.
(88) Baden, L. R.; El Sahly, H. M.; Essink, B.; Kotloff, K.; Frey, S.; Novak, R.; Diemert, D.; Spector, S. A.; Rouphael, N.; Creech, C. B.; McGettigan, J.; Khetan, S.; Segall, N.; Solis, J.; Brosz, A.; Fierro, C.; Schwartz, H.; Neuzil, K.; Corey, L.; Gilbert, P.; Janes, H.; Follmann, D.; Marovich, M.; Mascola, J.; Polakowski, L.; Ledgerwood, J.; Graham, B. S.; Bennett, H.; Pajon, R.; Knightly, C.; Leav, B.; Deng, W.; Zhou, H.; Han, S.; Ivarsson, M.; Miller, J.; Zaks, T. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* **2021**, *384* (5), 403–416. https://doi.org/10.1056/NEJMoa2035389.
(89) Berry, D. A. Bayesian Clinical Trials. *Nat. Rev. Drug Discov.* **2006**, *5* (1), 27–36. https://doi.org/10.1038/nrd1927.
(90) Hespanhol, L.; Vallio, C. S.; Costa, L. M.; Saragiotto, B. T. Understanding and Interpreting Confidence and Credible Intervals around Effect Estimates. *Braz. J. Phys. Ther.* **2019**, *23* (4), 290–301. https://doi.org/10.1016/j.bjpt.2018.12.006.
(91) James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. E. *An Introduction to Statistical Learning: With Applications in Python*; Springer texts in statistics; Springer: Cham, Switzerland, 2023.
(92) Bishop, C. M. *Pattern Recognition and Machine Learning*; Information science and statistics; Springer: New York, 2006.

(93) Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, Third edition.; Data science / machine learning; O'Reilly: Beijing Boston Farnham Sebastopol Tokyo, 2023.
(94) Murphy, K. P. *Probabilistic Machine Learning: An Introduction*; Adaptive computation and machine learning; The MIT Press: Cambridge, Massachusetts London, England, 2022.
(95) Lamplot, J. D.; Rodeo, S. A.; Brophy, R. H. A Practical Guide for the Current Use of Biologic Therapies in Sports Medicine. *Am. J. Sports Med.* **2020**, *48* (2), 488–503. https://doi.org/10.1177/0363546519836090.
(96) Schoch, B. S.; Werner, B. C.; Shapiro, S. A.; Camp, C. L.; Chalmers, P. N.; Cancienne, J. M. Effect of Bone Marrow Aspirate Concentrate and Platelet-Rich Plasma Augmentation on the Rate of Revision Rotator Cuff Repair. *Orthop. J. Sports Med.* **2022**, *10* (11), 23259671221127004. https://doi.org/10.1177/23259671221127004.
(97) Neff, P.; Franklin, D. B.; Jones, D. L.; Lang, S. D.; Nadone, H. R.; Gilmer, B. B.; Guttmann, D. Transtendinous Rotator Cuff Tear Repair with Bone Marrow Aspirate Concentrate Dermal Allograft Augmentation. *Arthrosc. Tech.* **2021**, *10* (4), e975–e980. https://doi.org/10.1016/j.eats.2020.11.019.
(98) Peterson, L. J. ORAL AND MAXILLOFACIAL SURGERY. *ORAL Surg. ORAL Med. ORAL Pathol.* **1998**, *85* (6).
(99) Marx, R. E. Platelet-Rich Plasma: Evidence to Support Its Use. *J. Oral Maxillofac. Surg.* **2004**, *62* (4), 489–496. https://doi.org/10.1016/j.joms.2003.12.003.
(100) Youn, G. M.; Woodall, B. M.; Elena, N.; Shin, E. C.; Pathare, N.; McGahan, P. J.; Chen, J. L. Arthroscopic Bone Marrow Aspirate Concentrate Harvesting From the Intercondylar Notch of the Knee. *Arthrosc. Tech.* **2018**, *7* (11), e1173–e1176. https://doi.org/10.1016/j.eats.2018.07.016.
(101) Keeling, L. E.; Belk, J. W.; Kraeutler, M. J.; Kallner, A. C.; Lindsay, A.; McCarty, E. C.; Postma, W. F. Bone Marrow Aspirate Concentrate for the Treatment of Knee Osteoarthritis: A Systematic Review. *Am. J. Sports Med.* **2022**, *50* (8), 2315–2323. https://doi.org/10.1177/03635465211018837.
(102) Gobbi, A.; Scotti, C.; Karnatzikos, G.; Mudhigere, A.; Castro, M.; Peretti, G. M. One-Step Surgery with Multipotent Stem Cells and Hyaluronan-Based Scaffold for the Treatment of Full-Thickness Chondral Defects of the Knee in Patients Older than 45 Years. *Knee Surg. Sports Traumatol. Arthrosc.* **2017**, *25* (8), 2494–2501. https://doi.org/10.1007/s00167-016-3984-6.
(103) Hernigou, P.; Mathieu, G.; Poignard, A.; Manicom, O.; Beaujean, F.; Rouard, H. Percutaneous Autologous Bone-Marrow Grafting for Nonunions. **2006**, *88*.
(104) Dulic, O.; Rasovic, P.; Lalic, I.; Kecojevic, V.; Gavrilovic, G.; Abazovic, D.; Maric, D.; Miskulin, M.; Bumbasirevic, M. Bone Marrow Aspirate Concentrate versus Platelet Rich Plasma or Hyaluronic Acid for the Treatment of Knee Osteoarthritis. *Medicina (Mex.)* **2021**, *57* (11), 1193. https://doi.org/10.3390/medicina57111193.

(105) Centeno, C.; Sheinkop, M.; Dodson, E.; Stemper, I.; Williams, C.; Hyzy, M.; Ichim, T.; Freeman, M. A Specific Protocol of Autologous Bone Marrow Concentrate and Platelet Products versus Exercise Therapy for Symptomatic Knee Osteoarthritis: A Randomized Controlled Trial with 2 Year Follow-Up. *J. Transl. Med.* **2018**, *16* (1), 355. https://doi.org/10.1186/s12967-018-1736-8.
(106) Maruvada, S.; Madrazo-Ibarra, A.; Varacallo, M. Anatomy, Rotator Cuff. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2023.
(107) He, Z.; Zhou, J. Empirical Evaluation of a New Method for Calculating Signal-to-Noise Ratio for Microarray Data Analysis. *Appl. Environ. Microbiol.* **2008**, *74* (10), 2957–2966. https://doi.org/10.1128/AEM.02536-07.
(108) Carlsson, A.; Persson, O.; Ingvarsson, J.; Widegren, B.; Salford, L.; Borrebaeck, C. A. K.; Wingren, C. Plasma Proteome Profiling Reveals Biomarker Patterns Associated with Prognosis and Therapy Selection in Glioblastoma Multiforme Patients. *PROTEOMICS – Clin. Appl.* **2010**, *4* (6–7), 591–602. https://doi.org/10.1002/prca.200900173.
(109) HUANG, S.; CAI, N.; PACHECO, P. P.; NARANDES, S.; WANG, Y.; XU, W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics* **2017**, *15* (1), 41–51. https://doi.org/10.21873/cgp.20063.
(110) *SPSS Modeler Subscription*. https://www.ibm.com/docs/en/spss-modeler/saas?topic=node-svm-expert-options (accessed 2025-04-21).
(111) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.
(112) Liu, Q.; Chen, C.; Zhang, Y.; Hu, Z. Feature Selection for Support Vector Machines with RBF Kernel. *Artif Intell Rev* **2011**, *36*, 99–115. https://doi.org/10.1007/s10462-011-9205-2.
(113) Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification. *Natl. Taiwan Univ.* **2003**, *101*, 1396–1400.
(114) Rodríguez-Pérez, R.; Vogt, M.; Bajorath, J. Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction. *ACS Omega* **2017**, *2* (10), 6371–6379. https://doi.org/10.1021/acsomega.7b01079.
(115) Mara, U. T. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests.
(116) Zar, J. H. *Biostatistical Analysis*, 4. ed., internat. ed.; Prentice Hall International: Upper Saddle River, NJ, 1999.
(117) Sullivan, G. M.; Feinn, R. Using Effect Size—or Why the P Value Is Not Enough. *J. Grad. Med. Educ.* **2012**, *4* (3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1.
(118) Junying Zhang; Lee, R.; Wang, Y. J. Support Vector Machine Classifications for Microarray Expression Data Set. In *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003*;

IEEE Comput. Soc: Xi'an, China, 2003; pp 67–71. https://doi.org/10.1109/ICCIMA.2003.1238102.
(119) Zou, D.; Man, H. A Novel Perspective towards SVM Combined with Autoencoder. *J. Phys. Conf. Ser.* **2022**, *2347* (1), 012011. https://doi.org/10.1088/1742-6596/2347/1/012011.
(120) Chicco, D.; Jurman, G. The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification. *BioData Min.* **2023**, *16*, 4. https://doi.org/10.1186/s13040-023-00322-4.
(121) Charilaou, P.; Battat, R. Machine Learning Models and Over-Fitting Considerations. *World J. Gastroenterol.* **2022**, *28* (5), 605–607. https://doi.org/10.3748/wjg.v28.i5.605.
(122) Jung, Y.; Hu, J. A K-Fold Averaging Cross-Validation Procedure. *J. Nonparametric Stat.* **2015**, *27* (2), 167–179. https://doi.org/10.1080/10485252.2015.1010532.
(123) Zhang, Y.; Wang, S.; Ji, G.; Dong, Z. An MR Brain Images Classifier System via Particle Swarm Optimization and Kernel Support Vector Machine. *Sci. World J.* **2013**, *2013*, e130134. https://doi.org/10.1155/2013/130134.
(124) Pal, A.; Ranzato, F.; Urban, C.; Zanella, M. Abstract Interpretation-Based Feature Importance for SVMs. arXiv October 22, 2022. http://arxiv.org/abs/2210.12456 (accessed 2023-12-20).
(125) Muttigi, M.; Han, I.; Park, H.-K.; Park, H.; Lee, S.-H. Matrilin-3 Role in Cartilage Development and Osteoarthritis. *Int. J. Mol. Sci.* **2016**, *17* (4), 590. https://doi.org/10.3390/ijms17040590.
(126) Klatt, A. R.; Becker, A.-K. A.; Neacsu, C. D.; Paulsson, M.; Wagener, R. The Matrilins: Modulators of Extracellular Matrix Assembly. *Int. J. Biochem. Cell Biol.* **2011**, *43* (3), 320–330. https://doi.org/10.1016/j.biocel.2010.12.010.
(127) Jayasuriya, C. T.; Goldring, M. B.; Terek, R.; Chen, Q. Matrilin-3 Induction of IL-1 Receptor Antagonist Is Required for up-Regulating Collagen II and Aggrecan and down-Regulating ADAMTS-5 Gene Expression. *Arthritis Res. Ther.* **2012**, *14* (5), R197. https://doi.org/10.1186/ar4033.
(128) Pei, M.; Luo, J.; Chen, Q. Enhancing and Maintaining Chondrogenesis of Synovial Fibroblasts by Cartilage Extracellular Matrix Protein Matrilins. *Osteoarthritis Cartilage* **2008**, *16* (9), 1110–1117. https://doi.org/10.1016/j.joca.2007.12.011.
(129) Zhang, Y.; Chen, Q. Changes of Matrilin Forms during Endochondral Ossification. Molecular Basis of Oligomeric Assembly. *J. Biol. Chem.* **2000**, *275* (42), 32628–32634. https://doi.org/10.1074/jbc.M002594200.
(130) Volk, S. W.; Luvalle, P.; Leask, T.; Leboy, P. S. A BMP Responsive Transcriptional Region in the Chicken Type X Collagen Gene. *J. Bone Miner. Res.* **1998**, *13* (10), 1521–1529. https://doi.org/10.1359/jbmr.1998.13.10.1521.
(131) Dreier, R. Hypertrophic Differentiation of Chondrocytes in Osteoarthritis: The Developmental Aspect of Degenerative Joint Disorders. *Arthritis Res. Ther.* **2010**, *12* (5), 216. https://doi.org/10.1186/ar3117.

(132) Yang, X.; Trehan, S. K.; Guan, Y.; Sun, C.; Moore, D. C.; Jayasuriya, C. T.; Chen, Q. Matrilin-3 Inhibits Chondrocyte Hypertrophy as a Bone Morphogenetic Protein-2 Antagonist *. *J. Biol. Chem.* **2014**, *289* (50), 34768–34779. https://doi.org/10.1074/jbc.M114.583104.
(133) Klatt, A. R.; Klinger, G.; Paul-Klausch, B.; Kühn, G.; Renno, J. H.; Wagener, R.; Paulsson, M.; Schmidt, J.; Malchau, G.; Wielckens, K. Matrilin-3 Activates the Expression of Osteoarthritis-Associated Genes in Primary Human Chondrocytes. *FEBS Lett.* **2009**, *583* (22), 3611–3617. https://doi.org/10.1016/j.febslet.2009.10.035.
(134) Wada, N.; Kimura, I.; Tanaka, H.; Ide, H.; Nohno, T. Glycosylphosphatidylinositol-Anchored Cell Surface Proteins Regulate Position-Specific Cell Affinity in the Limb Bud. *Dev. Biol.* **1998**, *202* (2), 244–252. https://doi.org/10.1006/dbio.1998.9013.
(135) Arthur, A.; Gronthos, S. Eph-Ephrin Signaling Mediates Cross-Talk Within the Bone Microenvironment. *Front. Cell Dev. Biol.* **2021**, *9*, 598612. https://doi.org/10.3389/fcell.2021.598612.
(136) Jing, X.; Miyajima, M.; Sawada, T.; Chen, Q.; Iida, K.; Furushima, K.; Arai, D.; Chihara, K.; Sakaguchi, K. Crosstalk of Humoral and Cell-Cell Contact-Mediated Signals in Postnatal Body Growth. *Cell Rep.* **2012**, *2* (3), 652–665. https://doi.org/10.1016/j.celrep.2012.08.021.
(137) Kozhemyakina, E.; Lassar, A. B.; Zelzer, E. A Pathway to Bone: Signaling Molecules and Transcription Factors Involved in Chondrocyte Development and Maturation. *Development* **2015**, *142* (5), 817–831. https://doi.org/10.1242/dev.105536.
(138) Stanley, E. R.; Berg, K. L.; Einstein, D. B.; Lee, P. S. W.; Pixley, F. J.; Wang, Y.; Yeung, Y.-G. Biology and Action of Colony-Stimulating Factor-1. *Mol. Reprod. Dev.* **1997**, *46* (1), 4–10. https://doi.org/10.1002/(SICI)1098-2795(199701)46:1%253C4::AID-MRD2%253E3.0.CO;2-V.
(139) Inoue, K.; Qin, Y.; Xia, Y.; Han, J.; Yuan, R.; Sun, J.; Xu, R.; Jiang, J. X.; Greenblatt, M. B.; Zhao, B. Bone Marrow Adipoq-Lineage Progenitors Are a Major Cellular Source of M-CSF That Dominates Bone Marrow Macrophage Development, Osteoclastogenesis, and Bone Mass. *eLife* **2023**, *12*, e82118. https://doi.org/10.7554/eLife.82118.
(140) Calandra, T.; Roger, T. Macrophage Migration Inhibitory Factor: A Regulator of Innate Immunity. *Nat. Rev. Immunol.* **2003**, *3* (10), 791–800. https://doi.org/10.1038/nri1200.
(141) Xu, F.; Teitelbaum, S. L. Osteoclasts: New Insights. *Bone Res.* **2013**, *1* (1), 11–26. https://doi.org/10.4248/BR201301003.
(142) Tanaka, S.; Takahashi, N.; Udagawa, N.; Tamura, T.; Akatsu, T.; Stanley, E. R.; Kurokawa, T.; Suda, T. Macrophage Colony-Stimulating Factor Is Indispensable for Both Proliferation and Differentiation of Osteoclast Progenitors. *J. Clin. Invest.* **1993**, *91* (1), 257–263. https://doi.org/10.1172/JCI116179.

(143) Landmann, R.; Müller, B.; Zimmerli, W. CD14, New Aspects of Ligand and Signal Diversity. *Microbes Infect.* **2000**, *2* (3), 295–304. https://doi.org/10.1016/S1286-4579(00)00298-7.
(144) Zamani, F.; Zare Shahneh, F.; Aghebati-Maleki, L.; Baradaran, B. Induction of CD14 Expression and Differentiation to Monocytes or Mature Macrophages in Promyelocytic Cell Lines: New Approach. *Adv. Pharm. Bull.* **2013**, *3* (2), 329–332. https://doi.org/10.5681/apb.2013.053.
(145) Sørensen, M. G.; Henriksen, K.; Schaller, S.; Henriksen, D. B.; Nielsen, F. C.; Dziegiel, M. H.; Karsdal, M. A. Characterization of Osteoclasts Derived from CD14+ Monocytes Isolated from Peripheral Blood. *J. Bone Miner. Metab.* **2006**, *25* (1), 36–45. https://doi.org/10.1007/s00774-006-0725-9.
(146) Shi, C.; Pamer, E. G. Monocyte Recruitment during Infection and Inflammation. *Nat. Rev. Immunol.* **2011**, *11* (11), 762–774. https://doi.org/10.1038/nri3070.
(147) Weivoda, M. M.; Bradley, E. W. Macrophages and Bone Remodeling. *J. Bone Miner. Res.* **2023**, *38* (3), 359–369. https://doi.org/10.1002/jbmr.4773.
(148) Elgueta, R.; Benson, M. J.; De Vries, V. C.; Wasiuk, A.; Guo, Y.; Noelle, R. J. Molecular Mechanism and Function of CD40/CD40L Engagement in the Immune System. *Immunol. Rev.* **2009**, *229* (1), 152–172. https://doi.org/10.1111/j.1600-065X.2009.00782.x.
(149) Ziegler, C. G.; Van Sloun, R.; Gonzalez, S.; Whitney, K. E.; DePhillipo, N. N.; Kennedy, M. I.; Dornan, G. J.; Evans, T. A.; Huard, J.; LaPrade, R. F. Characterization of Growth Factors, Cytokines, and Chemokines in Bone Marrow Concentrate and Platelet-Rich Plasma: A Prospective Analysis. *Am. J. Sports Med.* **2019**, *47* (9), 2174–2187. https://doi.org/10.1177/0363546519832003.
(150) Kristiansen, M.; Graversen, J. H.; Jacobsen, C.; Sonne, O.; Hoffman, H.-J.; Law, S. K. A.; Moestrup, S. K. Identification of the Haemoglobin Scavenger Receptor. *Nature* **2001**, *409* (6817), 198–201. https://doi.org/10.1038/35051594.
(151) Hu, J. M.; Liu, K.; Liu, J. H.; Jiang, X. L.; Wang, X. L.; Chen, Y. Z.; Li, S. G.; Zou, H.; Pang, L. J.; Liu, C. X.; Cui, X. B.; Yang, L.; Zhao, J.; Shen, X. H.; Jiang, J. F.; Liang, W. H.; Yuan, X. L.; Li, F. CD163 as a Marker of M2 Macrophage, Contribute to Predict Aggressiveness and Prognosis of Kazakh Esophageal Squamous Cell Carcinoma. *Oncotarget* **2017**, *8* (13), 21526–21538. https://doi.org/10.18632/oncotarget.15630.
(152) Guo, L.; Akahori, H.; Harari, E.; Smith, S. L.; Polavarapu, R.; Karmali, V.; Otsuka, F.; Gannon, R. L.; Braumann, R. E.; Dickinson, M. H.; Gupta, A.; Jenkins, A. L.; Lipinski, M. J.; Kim, J.; Chhour, P.; Vries, P. S. de; Jinnouchi, H.; Kutys, R.; Mori, H.; Kutyna, M. D.; Torii, S.; Sakamoto, A.; Choi, C. U.; Cheng, Q.; Grove, M. L.; Sawan, M. A.; Zhang, Y.; Cao, Y.; Kolodgie, F. D.; Cormode, D. P.; Arking, D. E.; Boerwinkle, E.; Morrison, A. C.; Erdmann, J.; Sotoodehnia, N.; Virmani, R.; Finn, A. V. *CD163+ macrophages promote angiogenesis and*

vascular permeability accompanied by inflammation in atherosclerosis. https://doi.org/10.1172/JCI93025.
(153) Sundman, E. A.; Cole, B. J.; Karas, V.; Della Valle, C.; Tetreault, M. W.; Mohammed, H. O.; Fortier, L. A. The Anti-Inflammatory and Matrix Restorative Mechanisms of Platelet-Rich Plasma in Osteoarthritis. *Am. J. Sports Med.* **2014**, *42* (1), 35–41. https://doi.org/10.1177/0363546513507766.
(154) Chen, H.; Deere, M.; Hecht, J. T.; Lawler, J. Cartilage Oligomeric Matrix Protein Is a Calcium-Binding Protein, and a Mutation in Its Type 3 Repeats Causes Conformational Changes. *J. Biol. Chem.* **2000**, *275* (34), 26538–26544. https://doi.org/10.1074/jbc.M909780199.
(155) Chang, L.-R.; Anand, P.; Varacallo, M. Anatomy, Shoulder and Upper Limb, Glenohumeral Joint. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2024.
(156) Eschweiler, J.; Horn, N.; Rath, B.; Betsch, M.; Baroncini, A.; Tingart, M.; Migliorini, F. The Biomechanics of Cartilage—An Overview. *Life* **2021**, *11* (4), 302. https://doi.org/10.3390/life11040302.
(157) Schmitz, P.; Cornelius Neumann, C.; Neumann, C.; Nerlich, M.; Dendorfer, S. Biomechanical Analysis of Iliac Crest Loading Following Cortico-Cancellous Bone Harvesting. *J. Orthop. Surg.* **2018**, *13* (1), 108. https://doi.org/10.1186/s13018-018-0822-1.
(158) Balusani, P.; Shrivastava, S.; Pundkar, A.; Kale, P. Navigating the Therapeutic Landscape: A Comprehensive Review of Platelet-Rich Plasma and Bone Marrow Aspirate Concentrate in Knee Osteoarthritis. *Cureus* **2024**. https://doi.org/10.7759/cureus.54747.
(159) Goujon, E. Recherches Expérimentales Sur Les Propriétés Physiologiques de La Moelle Des Os. *J Anat Physiol* **1869**, *6*, 399–412.
(160) Saeedi, P.; Halabian, R.; Imani Fooladi, A. A. A Revealing Review of Mesenchymal Stem Cells Therapy, Clinical Perspectives and Modification Strategies. *Stem Cell Investig.* **2019**, *6*, 34. https://doi.org/10.21037/sci.2019.08.11.
(161) Cavallo, C.; Boffa, A.; De Girolamo, L.; Merli, G.; Kon, E.; Cattini, L.; Santo, E.; Grigolo, B.; Filardo, G. Bone Marrow Aspirate Concentrate Quality Is Affected by Age and Harvest Site. *Knee Surg. Sports Traumatol. Arthrosc.* **2023**, *31* (6), 2140–2151. https://doi.org/10.1007/s00167-022-07153-6.
(162) Kehl, D.; Generali, M.; Mallone, A.; Heller, M.; Uldry, A.-C.; Cheng, P.; Gantenbein, B.; Hoerstrup, S. P.; Weber, B. Proteomic Analysis of Human Mesenchymal Stromal Cell Secretomes: A Systematic Comparison of the Angiogenic Potential. *Npj Regen. Med.* **2019**, *4* (1), 8. https://doi.org/10.1038/s41536-019-0070-y.
(163) McBride, J. D.; Rodriguez-Menocal, L.; Guzman, W.; Khan, A.; Myer, C.; Liu, X.; Bhattacharya, S. K.; Badiavas, E. V. Proteomic Analysis of Bone Marrow-Derived Mesenchymal Stem Cell Extracellular Vesicles from Healthy Donors: Implications for Proliferation, Angiogenesis, Wnt Signaling, and the Basement

Membrane. *Stem Cell Res. Ther.* **2021**, *12* (1), 328. https://doi.org/10.1186/s13287-021-02405-7.
(164) Wang, D.; Park, J. S.; Chu, J. S. F.; Krakowski, A.; Luo, K.; Chen, D. J.; Li, S. Proteomic Profiling of Bone Marrow Mesenchymal Stem Cells upon Transforming Growth Factor B1 Stimulation. *J. Biol. Chem.* **2004**, *279* (42), 43725–43734. https://doi.org/10.1074/jbc.M407368200.
(165) Park, D.; Koh, H.-S.; Choi, Y.-H.; Park, I. Bone Marrow Aspirate Concentrate (BMAC) for Knee Osteoarthritis: A Narrative Review of Clinical Efficacy and Future Directions. March 4, 2025. https://doi.org/10.20944/preprints202503.0139.v1.
(166) CDC. *Current Cigarette Smoking Among Adults in the United States*. Smoking and Tobacco Use. https://www.cdc.gov/tobacco/php/data-statistics/adult-data-cigarettes/index.html (accessed 2024-12-12).
(167) Ramirez, J. M.; Ribeiro, R.; Soldatkina, O.; Moraes, A.; Ferreira, P. G.; Melé, M. Cigarette Smoking Drives Accelerated Aging across Human Tissues.
(168) Nyunoya, T.; Monick, M. M.; Klingelhutz, A. L.; Glaser, H.; Cagley, J. R.; Brown, C. O.; Matsumoto, E.; Aykin-Burns, N.; Spitz, D. R.; Oshima, J.; Hunninghake, G. W. Cigarette Smoke Induces Cellular Senescence via Werner's Syndrome Protein Down-Regulation. *Am. J. Respir. Crit. Care Med.* **2009**, *179* (4), 279–287. https://doi.org/10.1164/rccm.200802-320OC.
(169) Akiyama, H.; Meyer, J. S.; Mortel, K. F.; Yasuo Terayama; Thornby, J. I.; Konno, S. Normal Human Aging: Factors Contributing to Cerebral Atrophy. *J. Neurol. Sci.* **1997**, *152* (1), 39–49. https://doi.org/10.1016/S0022-510X(97)00141-X.
(170) Asami, S. Cigarette Smoking Induces an Increase in Oxidative DNA Damage, 8- Hydroxydeoxyguanosine, in a Central Site of the Human Lung. *Carcinogenesis* **1997**, *18* (9), 1763–1766. https://doi.org/10.1093/carcin/18.9.1763.
(171) Valavanidis, A.; Vlachogianni, T.; Fiotakis, K. Tobacco Smoke: Involvement of Reactive Oxygen Species and Stable Free Radicals in Mechanisms of Oxidative Damage, Carcinogenesis and Synergistic Effects with Other Respirable Particles. *Int. J. Environ. Res. Public. Health* **2009**, *6* (2), 445–462. https://doi.org/10.3390/ijerph6020445.
(172) Seyler, T.; Mazumder, S.; Ahamed, R.; Zhu, W.; Blount, B. C.; Apelberg, B. J.; Wang, L. Tobacco Smoke Is a Major Source of Aromatic Amine Exposure in U.S. Adults: 2013–2014 National Health and Nutrition Examination Survey (NHANES). *Cancer Epidemiol. Biomarkers Prev.* **2023**, *32* (7), 986–994. https://doi.org/10.1158/1055-9965.EPI-23-0071.
(173) Ma, L.; Zheng, L. W.; Sham, M. H.; Cheung, L. K. Uncoupled Angiogenesis and Osteogenesis in Nicotine-Compromised Bone Healing. *J. Bone Miner. Res.* **2010**, *25* (6), 1305–1313. https://doi.org/10.1002/jbmr.19.
(174) El-Zawawy, H. B.; Gill, C. S.; Wright, R. W.; Sandell, L. J. Smoking Delays Chondrogenesis in a Mouse Model of Closed Tibial Fracture Healing. *J. Orthop. Res.* **2006**, *24* (12), 2150–2158. https://doi.org/10.1002/jor.20263.

(175) Wahl, E. A.; Schenck, T. L.; Machens, H.-G.; Egaña, J. T. Acute Stimulation of Mesenchymal Stem Cells with Cigarette Smoke Extract Affects Their Migration, Differentiation and Paracrine Potential. *Sci. Rep.* **2016**, *6* (1), 22957. https://doi.org/10.1038/srep22957.
(176) Tura-Ceide, O.; Lobo, B.; Paul, T.; Puig-Pey, R.; Coll-Bonfill, N.; García-Lucio, J.; Smolders, V.; Blanco, I.; Barberà, J. A.; Peinado, V. I. Cigarette Smoke Challenges Bone Marrow Mesenchymal Stem Cell Capacities in Guinea Pig. *Respir. Res.* **2017**, *18* (1), 50. https://doi.org/10.1186/s12931-017-0530-0.
(177) Holton, J.; Imam, M.; Ward, J.; Snow, M. The Basic Science of Bone Marrow Aspirate Concentrate in Chondral Injuries. *Orthop. Rev.* **2016**, *8* (3), 6659. https://doi.org/10.4081/or.2016.6659.
(178) Katagiri, T.; Watabe, T. Bone Morphogenetic Proteins. *Cold Spring Harb. Perspect. Biol.* **2016**, *8* (6), a021899. https://doi.org/10.1101/cshperspect.a021899.
(179) Yun, Y.-R.; Won, J. E.; Jeon, E.; Lee, S.; Kang, W.; Jo, H.; Jang, J.-H.; Shin, U. S.; Kim, H.-W. Fibroblast Growth Factors: Biology, Function, and Application for Tissue Regeneration. *J. Tissue Eng.* **2010**, *2010*, 218142. https://doi.org/10.4061/2010/218142.
(180) Hedbom, E.; Antonsson, P.; Hjerpe, A.; Aeschlimann, D.; Paulsson, M.; Rosa-Pimentel, E.; Sommarin, Y.; Wendel, M.; Oldberg, A.; Heinegård, D. Cartilage Matrix Proteins. An Acidic Oligomeric Protein (COMP) Detected Only in Cartilage. *J. Biol. Chem.* **1992**, *267* (9), 6132–6136. https://doi.org/10.1016/S0021-9258(18)42671-3.
(181) Chen, F. H.; Thomas, A. O.; Hecht, J. T.; Goldring, M. B.; Lawler, J. Cartilage Oligomeric Matrix Protein/Thrombospondin 5 Supports Chondrocyte Attachment through Interaction with Integrins. *J. Biol. Chem.* **2005**, *280* (38), 32655–32661. https://doi.org/10.1074/jbc.M504778200.
(182) Chen, T.; Ehnert, S.; Tendulkar, G.; Zhu, S.; Arnscheidt, C.; Aspera-Werz, R. H.; Nussler, A. K. Primary Human Chondrocytes Affected by Cigarette Smoke—Therapeutic Challenges. *Int. J. Mol. Sci.* **2020**, *21* (5), 1901. https://doi.org/10.3390/ijms21051901.
(183) Baumgarten, K. M.; Gerlach, D.; Galatz, L. M.; Teefey, S. A.; Middleton, W. D.; Ditsios, K.; Yamaguchi, K. Cigarette Smoking Increases the Risk for Rotator Cuff Tears. *Clin. Orthop. Relat. Res.* **2010**, *468* (6), 1534. https://doi.org/10.1007/s11999-009-0781-2.
(184) Naimark, M.; Robbins, C. B.; Gagnier, J. J.; Landfair, G.; Carpenter, J.; Bedi, A.; Miller, B. S. Impact of Smoking on Patient Outcomes after Arthroscopic Rotator Cuff Repair. *BMJ Open Sport Exerc. Med.* **2018**, *4* (1). https://doi.org/10.1136/bmjsem-2018-000416.
(185) Carbone, S.; Gumina, S.; Arceri, V.; Campagna, V.; Fagnani, C.; Postacchini, F. The Impact of Preoperative Smoking Habit on Rotator Cuff Tear: Cigarette Smoking Influences Rotator Cuff Tear Sizes. *J. Shoulder Elbow Surg.* **2012**, *21* (1), 56–60. https://doi.org/10.1016/j.jse.2011.01.039.

(186) Zabrzyński, J.; Huri, G.; Gagat, M.; Łapaj, Ł.; Yataganbaba, A.; Szwedowski, D.; Askin, M.; Paczesny, Ł. The Impact of Smoking on Clinical Results Following the Rotator Cuff and Biceps Tendon Complex Arthroscopic Surgery. *J. Clin. Med.* **2021**, *10* (4), 599. https://doi.org/10.3390/jcm10040599.
(187) Osti, L.; Buda, M.; Del Buono, A. Fatty Infiltration of the Shoulder: Diagnosis and Reversibility. *Muscle Ligaments Tendons J.* **2019**, *03* (04), 351. https://doi.org/10.32098/mltj.04.2013.18.
(188) Burkhart, S. S.; Barth, J. R. H.; Richards, D. P.; Zlatkin, M. B.; Larsen, M. Arthroscopic Repair of Massive Rotator Cuff Tears With Stage 3 and 4 Fatty Degeneration. *Arthrosc. J. Arthrosc. Relat. Surg.* **2007**, *23* (4), 347–354. https://doi.org/10.1016/j.arthro.2006.12.012.
(189) Caruso, R.; O'Connor, R.; Stephens, W.; Cummings, K.; Fong, G. Toxic Metal Concentrations in Cigarettes Obtained from U.S. Smokers in 2009: Results from the International Tobacco Control (ITC) United States Survey Cohort. *Int. J. Environ. Res. Public. Health* **2013**, *11* (1), 202–217. https://doi.org/10.3390/ijerph110100202.

Vita

Colin Herna was born on June 3, 1998, in Bethlehem, Pennsylvania. He is the son of Glenn and Gail and brother of Austin and Jonathan. He earned a B.S. in Biochemistry and Molecular Biology from DeSales University in May 2020. Throughout his time as an undergraduate student, Colin was fortunate enough to work on several research projects under Dr. Sarah Hayik, Dr. Steven Sweeney, and Dr. Julie Himmelberger. With their guidance and support, Colin opted to pursue his doctoral degree at Lehigh University in the Bioengineering program under the mentorship of Dr. Sabrina Jedlicka. Colin's work focused on an autologous, bio-injectable treatment called bone marrow aspirate concentrate (BMAC). The aim was to characterize the proteins found in BMAC by leveraging data science techniques such as machine learning, data visualization, and statistical pipelines. Colin's plan is to break into the quantitative finance world or the pharmaceutical industry with the skills and experience he has acquired throughout his academic career.